Universidad
Carlos III de Madrid
www.uc3m.es

Trabajo de Fin de Grado

Modelo Matemático de Predicción de Riesgos Bancarios

Nicolás Baviano Pérez-Tabernero
Grado en Ingeniería en Tecnologías Industriales
Tutora: M. Carmen Aguilera Morillo
Departamento de Estadística
Madrid, Mayo de 2015

"Scientia potestas est "
Francis Bacon, 1597

Agradecimientos

Este trabajo no habría sido posible de no ser por la implicación de mi tutora, Carmen Aguilera, el espíritu inconformista de Javier Calvo y el carácter infatigable de Javier Rebanal. Muchas gracias.

Resumen

A diario, las redes sociales generan cantidades masivas de información. El objetivo de este Trabajo de Fin de Grado (TFG) es analizar la calidad de los datos disponibles en dichas redes sociales mediante un ejercicio cuantitativo. Para ello se construirá un modelo matemático de predicción de mora con variables obtenidas en dichas redes sociales. Este trabajo abarca desde la obtención de datos hasta su inclusión en un modelo predictivo.

Se partirá de una base de datos real cedida por una entidad financiera, de la que se obtendrá la probabilidad de mora de algunos clientes, y sobre la cual se desea construir el nuevo modelo.

El trabajo se divide en 7 secciones. La primera de ellas sirve de justificación del trabajo llevado a cabo. La segunda consta de un estudio sobre los modelos predictivos y su relación con el sector financiero. En las secciones tercera y cuarta se detalla la obtención y el tratamiento de datos respectivamente. En la quinta sección se expone la metodología de construcción de un modelo de regresión logística, dando paso en la sexta sección a la construcción de un modelo con los datos previamente descritos. Finalmente, se reserva una sección para analizar las conclusiones del modelo construido y las futuras líneas de investigación.

Abstract

Daily, social networks create massive data. The aim of this Bachelor Thesis is to analyze the quality of the data available in these social networks through a quantitative exercise. Therefore a predictive mathematical model will be built including social networks variables. This Bachelor Thesis includes from the process of data obtaining to their inclusion in a predictive model.

The starting point will be a real database ceded from a financial institution, from which we will obtain some clients probability of default, and over which the new model will be built.

The Thesis has 7 sections. The first one will support the being of this thesis. The second section will consist of a study on the predictive models and its usage in the finance sector. In sections third and fourth data obtaining and transformation will be described. In the fifth section the methodology of the logistic regression model is presented, stepping, in the sixth section, into the model building, including social network's data. Finally, a seventh section is kept for analyzing the conclusions of the built model and the path for future investigations.

Índice de contenidos

Índice de figuras

Índice de tablas

1. Introducción

1.1 Motivación

El presente Trabajo de Fin de Grado se presenta como un ensayo de investigación sobre un tema en alza: el análisis de grandes cantidades de datos. Surge ante la idea de que todos los datos que nos rodean son útiles en su medida, y ante la necesidad de acompasar la evolución de la metodología en el sector financiero con el avance de las tecnologías de la información. La motivación de este trabajo se origina durante la realización de prácticas en la empresa Management Solutions.

1.2 Objetivos

El objetivo de este Trabajo de Fin de Grado no es nada desdeñoso. A tenor de las corrientes de pensamiento que ponen en alza el valor de los datos contenidos en redes sociales, se trata de construir un modelo de scoring a partir de información extraída de dichas redes. Se pretende evaluar la calidad de la información disponible en la web, y proponer una aplicación en el sector financiero.

1.3 Metodología y retos

Se empleará metodología típica en el ámbito de la estadística, la programación y la gestión de bases de datos, conjugado con un conocimiento del sector financiero que permitirá el desarrollo y, en caso de éxito, implantación del nuevo modelo. Se trata de un ejercicio que no ha sido planteado hasta ahora, por lo que durante el desarrollo del Trabajo surgirán impedimentos con los que no se contaba en un principio. Además, al ser un proyecto innovador, no existe apenas literatura sobre aplicaciones similares al tema tratado.

2. Modelos predictivos en el sector financiero

Un modelo matemático es una representación analítica de todo o parte de un sistema real. Los modelos ayudan a comprender el funcionamiento de dichos sistemas, estudiar los efectos de los distintos componentes o predecir posibles comportamientos bajo ciertos cambios. Además, ofrecen información objetiva automática y eficiente.

Los modelos matemáticos se pueden clasificar según la información de entrada, el tipo de representación, su aleatoriedad y según su objetivo. En nuestro caso, trataremos con modelos predictivos: modelos en los cuales no obtenemos una salida específica sino la probabilidad de dicha salida. Existe por tanto una incertidumbre en dichos modelos que será evaluada.

Los modelos predictivos juegan un papel clave en muchas áreas de conocimiento. Por citar algunos ejemplos, en medicina, diferentes modelos permiten predecir la probabilidad de contraer una enfermedad; en meteorología predecir el tiempo o simular fenómenos atmosféricos; y en Marketing predecir comportamientos de los clientes.

2.1 Introducción

En concreto, en el sector financiero el uso de modelos predictivos ha crecido considerablemente, de modo que hoy en día una elevada proporción de las decisiones de las entidades están automatizadas mediante modelos de decisión (MIT, 2005).

Una primera división clara que podemos realizar entre los modelos son aquellos orientados a la gestión comercial y aquellos orientados a la gestión del riesgo.

Sobre los modelos con orientación comercial, comentar que tienen como centro de estudio al cliente, ya sea por su captación, estimar su nivel de gasto, probabilidad de abandono o predecir la propensión de compra de nuevos productos, entre otros. Emplean técnicas estadísticas diversas como las regresiones lineales, modelos Logit y Probit, redes bayesianas, clustering o árboles de decisión.

La gestión del riesgo toma una importancia vital en el sector financiero. Es por ello que, por el acuerdo de Basilea III, todas las entidades están obligadas a aprovisionarse en función del riesgo de su cartera, lo cual incide directamente sobre la cuenta de resultados y la estrategia corporativa del banco. Entre los modelos predictivos orientados a la gestión del riesgo, encontramos modelos destinados a estimar la probabilidad de impago (*Probability of Default, PD*), cuantificar la pérdida por impago (*Loss Given Default, LGD*) o la exposición al riesgo de crédito. Dado el amplio abanico de modelos existentes, centraremos nuestra atención en los modelos de calificación (Sección 2.2).

Para el desarrollo de los modelos predictivos existe una gran variedad de entornos estadísticos, de los cuales se describen los más comunes en la Tabla 1.

	Descripción	Recomendado para las siguientes técnicas
	Excel es una aplicación basada en hojas de cálculo. Su uso para modelización es muy manual, pero muy sencillo para usuarios con poco conocimientos en programación	Cadenas de Markov, Monte Carlo, árboles de clasificación y regresión
	SAS es un paquete estadístico que ejecutiva distintos procesos a través de una serie de comandos. El programa básico permite ejecutar una gran cantidad de procesos distintos, pero para algunos es necesario adquirir software	Modelos logit y probit, clustering, mapas autoorganizativos, redes neuronales, cadenas de Markov…
	R es un entorno libre de análisis estadístico de datos y de creación de gráficos estadísticos. Se basa en una interfaz de usuario de líneas de comandos	Series temporales, redes neuronales, cadenas de Markov, optimización lineal y no lineal
	SPSS es un sistema de análisis estadístico y gestión de información. Su aplicación fundamental está orientada al análisis multivariante de datos experimentales. Su uso es a través de ventanas y menús	Series temporales, optimización y clustering
	Python es un lenguaje de programación que permite realizar análisis estadístico de datos de manera eficiente. Recientemente ha adquirido mucha relevancia	Series temporales, redes neuronales, cadenas de Markov, optimización lineal y no lineal

Tabla 1 Entornos estadísticos más utilizados en el sector financiero

2.2 Modelos de calificación

Los modelos de calificación tienen por objeto clasificar clientes u operaciones en grupos homogéneos a efectos de Riesgo de Crédito. Evalúan la solvencia y la calidad crediticia del cliente en base a información profesional, demográfica y personal, además de anteriores operaciones (en caso de existir). En cuanto a la metodología, utilizan modelos con perfil probabilístico como los Logit o Probit. A continuación se hace una relación detalla de los modelos de calificación:

- Según el objeto de calificación:
  - **Modelos de Rating:** clasifican a los clientes por su probabilidad de impago. Se utilizan en carteras mayoristas (empresas, entidades financieras), dado que se considera que el cliente se va a comportar de manera homogénea en todas sus operaciones. Esto es, si una empresa entrara en situación de impago no cumplirá ninguna de sus deudas, ya sea una tarjeta de crédito o la factura de la luz. En ocasiones se emplean redes neuronales para el desarrollo de este tipo de modelos.
  - **Modelos de Scoring:** clasifican cada operación asociada a cada cliente, como una hipoteca, un crédito al consumo o una tarjeta de crédito. Se aplican por tanto a carteras de personas físicas y microempresas. Son muy similares a los modelos de Rating, con la salvedad de que evalúan cada operación por separado, ya que el cliente se comporta de forma distinta con cada obligación.
- Según la finalidad del modelo:
  - **Modelos de admisión:** Se emplean ante la solicitud de una nueva operación. Al igual que los modelos de Scoring, califican el binomio Cliente-Operación. Asigna una probabilidad de incumplimiento al cliente en el momento de la solicitud utilizando variables conocidas hasta ese momento.
  - **Modelos de seguimiento:** Se emplean periódicamente para reevaluar la calidad crediticia de la cartera. Asigna una probabilidad de

incumplimiento en un momento distinto al de su solicitud, por lo que puede emplear variables de comportamiento del contrato. Se realizan modelos de seguimiento de la operación y del cliente.
  - **Modelos proactivos:** Se apoyan en los modelos de seguimiento y se utilizan bajo petición para identificar a qué parte de una cartera se le puede preconceder una operación. Además determinan el límite por producto y por cliente.
- Según la información empleada:
  - **Modelos de incumplimiento:** Utilizan información histórica de operaciones o clientes morosos de la entidad para inferir el patrón de morosidad.
  - **Modelos de réplica:** Ante la inexistencia de información interna de incumplimientos, extrapolan el patrón de morosidad a partir de la calificación otorgada por agencias externas (S&P, Moody's, Fitch).
  - **Modelos Expertos:** Son los modelos internos en los cuales el ejecutor es quién toma la decisión basándose en variables cualitativas. Estos modelos normalmente complementan con información cualitativa los modelos cuantitativos comentados anteriormente.

En nuestro caso de estudio, se parte de una base de datos construida para un modelo de Scoring. Por lo tanto, aunque la realización de un modelo de seguimiento o de rating podría ser interesante, se considera la realización de un modelo de Scoring construido con variables extraídas de redes sociales.

2.3 Minería de datos en el sector financiero

El sector financiero tiene una estrecha relación con la Minería de datos. Las entidades financieras producen cantidades de datos colosales diariamente, lo que motiva la aplicación de técnicas de análisis de datos. Se confeccionan grandes bases de datos con información centrada en grupos de variables básicas: datos referentes a su relación con el banco, es decir, qué productos tiene y cómo ha sido su trayectoria interna en la entidad; variables socioeconómicas y demográficas del cliente, como edad, situación laboral, familiar, etc.; datos de calidad de riesgo, que se refieren a su historial de pagos y, por último, variables de operación, como cantidad solicitada en caso de solicitar un crédito, plazos de amortización, etc.

Se podría decir que el sector financiero tiene dos acercamientos a la Minería de Datos. Por un lado, análisis de inversiones, econometría, comercio o análisis largoplacistas. Este ámbito no es novedoso ya que se lleva practicando tiempo.

Sin embargo, tiene un acercamiento más reciente dado que la minería de datos permite realizar un análisis con mayor trascendencia, como por ejemplo un conocimiento más profundo de los comportamientos de los clientes, del desempeño de la entidad o de la competencia de la misma (Magoulas, 2011).

Aunque el crecimiento del Big Data viene propiciado por la caída en los costes de creación y almacenamiento de datos, y el desarrollo de plataformas que permiten tratar esa cantidad de información, lo que realmente motiva el interés del Big Data es la explosión de tipos de datos procesados. Gracias a las técnicas de procesamiento de lenguajes naturales y aprendizaje automático, es posible

modelizar datos cualitativos, cosa que sin dichas técnicas resulta tremendamente complicado cuando no imposible. Un ejemplo llamativo al respecto fue la predicción de los precios de la bolsa de Nueva York analizando los mensajes de Twitter (Zhang, 2013).

En el sector financiero se han empleado prácticamente todas las técnicas estadísticas para el análisis de datos: Simulación de Monte Carlo para predecir el precio de las opciones (Huang, 2004), redes neuronales (Kovalerchuk, 2009), clustering (Walzack, 2001)o regresiones lineales entre otros. Algunas de las más sofisticadas, como una fusión de técnicas de redes neuronales con algoritmos genéticos (Loofbourrow, 1995) tampoco se han consagrado como técnicas definitivas, sino que abren otra puerta para la investigación. También se usan ampliamente técnicas de evaluación como el *bootstrapping* (Cogneau & Zakamouline, 2010) .

La minería de datos ofrece además cierta protección contra los riesgos que asume una entidad financiera. Sumado a esto, la minería de datos permite detectar fraudes e irregularidades en tiempo real, lo que reduce significativamente la pérdida tanto monetaria como de imagen o de clientes. Como era de esperar, la minería de datos ha tomado especial relevancia con la última crisis financiera, que puso en evidencia la debilidad de los modelos financieros, las teorías sobre precios y fondos de inversión o la incorrecta gestión y cuantificación de riesgos.

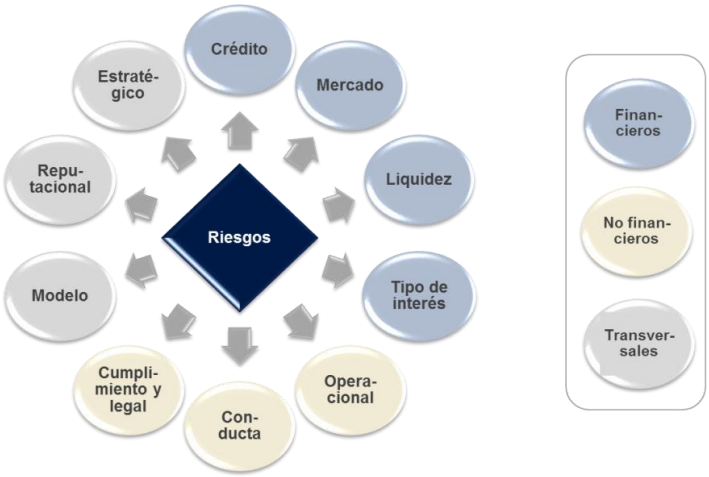

Figura 1 Mapa ilustrativo de riesgos. Fuente: (Management Solutions, 2014)

2.4 El riesgo de modelo

En esta sección se tratarán los riesgos que se asumen cuando se utiliza un modelo. Como se ha comentado anteriormente, existe poca normativa referente al riesgo de modelo. La gestión del riesgo se ha comenzado a regularizar desde la Reserva Federal de EEUU (Fed), pero existe aún poca normativa referente al riesgo de modelo. Se trata de la *Supervisory Guidance on Model Risk Management*, (OCC, 2011-12) donde se establecen unas directrices con buenas prácticas en la construcción de modelos, que abarcan desde el desarrollo de un modelo hasta su implantación, incluido el control y la documentación por parte de todos los intervinientes.

En cualquier caso, es el Consejo de Administración de la entidad el último responsable de los riesgos de modelo, y debe ser informado periódicamente sobre los riesgos a los que pudiera estar expuesta la entidad.

El riesgo de modelo se define como <<el conjunto de posibles consecuencias adversas derivadas de decisiones basadas en resultados e informes incorrectos de modelos, o de su uso inapropiado>> (Fed, 2012). El riesgo de modelo está presente durante toda la vida del modelo, desde su concepción hasta su auditoría, pasando por la aplicación, seguimiento y validación. Los errores pueden provenir de tres aspectos (Management Solutions, 2014):

1. Carencias en los datos, tanto de disponibilidad como de calidad, incluyendo errores en los datos, falta de profundidad histórica o fallos en la alimentación de las variables. Pueden ocurrir tanto en la creación como en la alimentación del modelo.
2. Incertidumbre en la estimación o errores en el modelo, en la forma de las simplificaciones, aproximaciones, hipótesis erróneas o creación desacertada del modelo. Pueden ocurrir desde su creación hasta su implantación.
3. Uso inadecuado del modelo, que incluye tanto su aplicación fuera del uso para el que fue diseñado como el hecho de no recalibrar el modelo de manera periódica.

2.4.1 Carencias en los datos

La obtención de datos en ocasiones puede suponer un reto. Si la entidad financiera no tiene una base de datos actualizada y bien estructurada para todos los departamentos de la empresa, los datos no son fácilmente accesibles. Por ello, a la hora de construir un modelo debemos considerar la cantidad y la calidad de los datos (puede ser que estén desactualizados). Se debe analizar la base de datos por si faltara alguna variable crítica en algún registro, ya que sin dichas variables la capacidad de predicción del modelo se ve seriamente expuesta.

Para cuantificar dichos errores, se puede realizar un análisis de la salida del modelo en ausencia de variables críticas, desprendiéndose una idea de la destreza del modelo.

A la hora de mitigar este tipo de error se debe realizar un análisis exhaustivo de la calidad de los datos (*data quality*). Se estima que los errores debidos al *data quality* le cuestan a la economía estadounidense 600.000 millones de dólares (Eckerson, 2002).

2.4.2 Incertidumbre en las estimaciones

Por definición, un modelo es una simplificación de la realidad. En dicha simplificación puede haber errores causados por la incertidumbre implícita en los estimadores. Una mala praxis en la creación del modelo puede llevar a la pérdida de predicción del mismo.

Para mitigar este error se puede realizar un *back test* del modelo, que coteje la salida prevista del modelo con la observada. También se debe realizar un *stress test* que someta al modelo a diferentes estados de tensión en la alimentación del mismo y analizar el desempeño en esas situaciones.

2.4.3 Uso inadecuado del modelo

Los modelos se conciben para usos muy concretos, siendo inservibles en caso de ser utilizado para otro fin o en otro ámbito. Por ejemplo, no se podría utilizar el mismo modelo de crédito en España y en Francia, como tampoco se puede emplear el mismo modelo para la concesión de hipotecas o créditos. El modelo sólo será altamente predictivo si su uso es adecuado.

Para mitigar este error se debe realizar un seguimiento del modelo, actualizándolo regularmente, incluyendo la monitorización automatizada de su poder predictivo o un sistema de alertas tempranas de deterioro. Además de lo anterior, es conveniente una correcta documentación del modelo, que permita la réplica por parte de terceros así como el traspaso a un nuevo modelizador sin pérdida de conocimiento.

3. Fuentes de datos

En esta sección se detallarán las fuentes de datos empleadas en este Trabajo. Se estudiarán las posibles fuentes de datos para realizar el modelo, evaluando las redes sociales más populares en España hasta la fecha, así como la información que podemos extraer de ellas y la relevancia de dicha información. Además, se describirá la base de datos cedida por una entidad financiera así como las consideraciones a tener al tratar con una base de datos similar.

3.1 Redes sociales

3.1.1 Las redes sociales en España

En España, un 79% por cierto de la población utiliza redes sociales. Se encuentra entre los 10 primeros países de la UE en cuanto a uso de las redes sociales y el tiempo que les dedicamos. Son los jóvenes entre 18 y 30 años los que mayor uso hacen de estas redes.

Según un estudio realizado por la agencia IAB, en España el 99% de la población conoce Facebook, y un 94% lo utiliza. Le siguen en cuanto a uso Twitter, YouTube, Tuenti, Google+, Instagram y LinkedIn. Estudiaremos los datos que nos ofrecen éstas redes sociales así como su viabilidad para la extracción de éstos.

El uso que se realiza de las redes no se limita a lo estrictamente social, sino que las empresas también sacan provecho de ellas. Se apoyan en las redes para ofrecer promociones, publicar ofertas de empleo, organizar concursos o dar información sobre un producto entre otras aplicaciones. Por poner un ejemplo, el 93% de los usuarios de Facebook sigue alguna marca: esa marca tiene una vía de comunicación abierta con ese cliente (IAB). Si quisiera lanzar una promoción a unos clientes con unas características determinadas (edad, sexo, ubicación…) podría hacerlo desde dicha plataforma.

Según un informe de la Asociación Española de la Economía Digital (adigital)el 85% de las empresas utilizan estas plataformas con fines de negocio. Facebook y Twitter son las dos redes sociales a las que más recurren las empresas encuestadas que buscan esos propósitos, con un 79,29% y un 79,44%, respectivamente. La tercera posición la ocupa LinkedIn, con un 51,48% (adigital).

Es por ello que cada día toma más importancia el análisis de las redes sociales, pues la mitad de las empresas afirman que la inversión realizada en campañas en redes sociales es superada por el retorno que obtienen de dichas campañas. En la figura 2 se muestra en porcentaje el conocimiento que se tiene de las principales redes sociales.

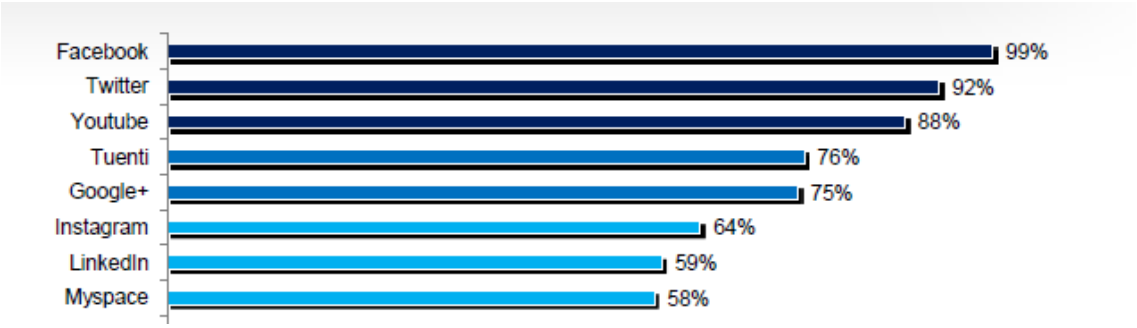

Figura 2 Porcentaje de encuestados que afirma conocer determinada red social. Fuente: (IAB)

3.1.2 Facebook

Se trata de la red social más extendida, con 1320 millones de usuarios registrados en Octubre de 2014 (fuente: Wikipedia) alrededor de todo el mundo. Un usuario medio en España dedica entre 4 y 5 horas a la semana a esta red, con una penetración del 94% de la población. Inicialmente se creó para facilitar el contacto entre estudiantes, pero recientemente el mundo empresarial ha comenzado a aprovechar Facebook para potenciar sus actividades. Los usuarios de Facebook ponen a disposición pública gran cantidad de datos, compartiendo información personal, demográfica, ubicación o estado civil, que podrían ser variables de interés para el ámbito financiero.

Para la obtención de datos de Facebook existen diferentes alternativas.

Una de ellas consiste en utilizar la Interfaz de Programación de Aplicaciones (*API* por sus siglas en inglés) de Facebook a través de la cual es posible desarrollar un programa que permita obtener datos para realizar un estudio. Facebook pone a disposición del usuario una serie de herramientas para explotar la red social y poder aprovechar sus datos. Se podría programar un programa en Python que, utilizando la API de Facebook, realiza una consulta entre los datos de las redes sociales, extrayendo así la información que se desee. Por poner un ejemplo de aplicación, dentro de la página de desarrolladores de Facebook existe un apartado exclusivamente para el comercio electrónico. Contiene herramientas que permiten aprovechar la basta fuente de información que supone Facebook para mejorar dicho negocio.

La principal contraindicación de utilizar una API es que tienen las funcionalidades limitadas, y para realizar un estudio exhaustivo es posible que sea necesario contactar con Facebook para ampliar los permisos.

Por otro lado, existen gran cantidad de herramientas informáticas dedicadas a la extracción de datos de Facebook. Utilizan métodos de dudosa legalidad por lo que lo dejaremos fuera del estudio.

Una opción de elevado interés sería crear un perfil y pedir a nuestros clientes que añada como amigo a su red dicho perfil. Como usuario de Facebook se pueden descargar los datos de tus contactos, e incluso realizar un análisis de dichos datos sin necesidad de almacenarlos. Se dispondría de una base de datos que se actualiza de manera autónoma y con datos limpios (No habría error en la descarga de datos). Algunas empresas ya aprovechan éste método pidiendo a sus clientes que se unan a su red en Facebook.

Aunque son alternativas con elevado interés, requieren un tiempo que no se dispone para la elaboración de este trabajo. Además la existencia de otras redes

sociales con variables de mayor interés para el sector financiero nos empuja a descartar esta opción.

3.1.3 Twitter

Twitter es una red social creada en primera instancia para su uso por periodistas. Se trata de una forma rápida de compartir información, aunque de manera limitada (140 caracteres/tweet). Con la proliferación de Twitter, hoy en día su uso se ha extendido a cualquier usuario. Un 49% de la población española utiliza Twitter dedicándole de media entre 3 y 4 horas a la semana. En Twitter encontramos por lo general mensajes subjetivos y poca o nula información de interés financiero.

No obstante, la extracción de datos de Twitter resulta aparentemente sencilla, ya que la información de Twitter es pública y hay cantidad de herramientas y paquetes que simplifican la extracción masiva de datos para realizar un análisis.

De la misma forma que Facebook, Twitter pone a disposición de los desarrolladores su API para la extracción de datos. Utilizando programas estadísticos como R es posible realizar un análisis de los datos de Twitter. Sin embargo, aquí es donde surgen las complicaciones en la minería de datos: la información que se puede descargar se compone cadenas de texto, y para poder tratar con esa información es necesario un Procesador de Lenguajes Naturales.

En el ámbito financiero destaca un trabajo de la Universidad de Texas en el cual, descargando datos masivamente de Twitter y realizando un *sentiment analysis* sobre los tweets (mensajes de Twitter), fueron capaces de predecir el comportamiento de los precios de la bolsa (ver (Zhang, 2013) para más detalle). Destacan también estudios de diferentes empresas, las cuales buscan entre los usuarios de Twitter a personas con un número elevado de seguidores, para ofrecerles ventajas o promociones a cambio de publicar mensajes en los que hablen de sus empresas.

En cualquier caso, nuestro estudio se centra en la aceptación o no de un cliente, por lo que estos métodos no resultan de interés para nuestro trabajo.

3.1.4 LinkedIn

LinkedIn es la red social profesional por excelencia. Supera los 300 millones de usuarios, y permite tener información del ámbito laboral de los contactos. Por lo general se comparte información sobre la trayectoria académica, experiencia laboral, idiomas que habla el usuario, intereses o aptitudes, entre otros. En España, un 22% de la población utiliza LinkedIn, dedicándole de media algo más de 2 horas semanales.

Se trata de la red social más interesante para llevar a cabo nuestro estudio, ya que las variables que podemos extraer de ella son de alto interés financiero, por ejemplo sector en el que se trabaja, años de experiencia, periodos en paro, formación, ubicación o publicaciones. Además se trata de una red social con información objetiva y, presuponemos, con alto nivel de veracidad, lo que hace de ella una alternativa muy prometedora.

Al igual que el resto de redes sociales citadas en este trabajo, LinkedIn pone a disposición del usuario su API para el tratamiento de datos. Sin embargo con dicha

API no es posible descargar información, sino que se puede analizar pero sin ser almacenada de forma permanente.

La privacidad de los datos de LinkedIn varía en función de las preferencias del usuario. Por defecto, un perfil puede ser visto por cualquier usuario de la red, y cualquier usuario puede conectar con otro. Sin embargo, los perfiles de LinkedIn tienen dos *interfaces*: una privada, la cual pueden ver los contactos y en la que se muestra toda la información que el usuario tenga colgada en la red, y otra pública, a la que puede acceder cualquier usuario de internet de manera anónima. No se necesita tener perfil en LinkedIn para obtener la información pública de un usuario. Es la información que se muestra buscando el nombre de una persona en LinkedIn a través de Google (escribiendo en Google: *Nombre Apellidos site: linkedin.com*)

En el perfil público se encuentra información acerca de la trayectoria profesional y académica, con sus correspondientes fechas, idiomas que habla el usuario, grupos, aptitudes y voluntariados. Estos datos son suficientes para llevar a cabo nuestro trabajo, ya que son variables de las que no dispone una entidad financiera y se presupone que podrían arrojar mucha información al estudio. Además, su fácil acceso hace posible la inclusión en el trabajo.

3.1.5 Otras redes sociales

Se han evaluado también otras redes sociales, como Google+, Tumblr, Pinterest o Instagram. Finalmente, se han descartado todas, ya que no ofrecen información con el interés que nos suscita LinkedIn, a pesar de que algunas de estas redes tienen una penetración mayor en el territorio español (IAB).

3.2 Extracción de datos de redes sociales con Microsoft VBA

3.2.1 Método empleado

En la presente sección se tratará la extracción de datos de una red social. Como se describe a continuación, el método empleado es algo lento. Por ello, será válido para muestras de estudio, pero para una descarga mayor de datos sería aconsejable utilizar otras herramientas más potentes (detalladas en la sección 3.1.3).

Es de importancia destacar que la obtención de datos realizada es para un estudio, y por tanto si el método que aquí se describe fuera llevado a cabo para otro fin habría que tener en cuenta consideraciones éticas y legales. Además, los Términos y Condiciones de Uso de algunas redes sociales desaconsejan la extracción de datos por diferentes motivos, pero principalmente por la Ley Orgánica de Protección de Datos.

Los datos para la muestra de nuestro estudio se han obtenido mediante un proceso semiautomático de búsqueda de información, desarrollado en una plataforma muy sencilla: Microsoft Visual Basic Applications para Office (VBA).

Excel VBA permite escribir programas, llamados macros, los cuales a su vez pueden hacer uso de otros programas instalados en el equipo. Por ejemplo, pueden mandar correos desde Microsoft Outlook, automatizar la creación de presentaciones con PowerPoint, o acceder a internet con Microsoft Internet Explorer. Será con esta última función con la que se automatizará el proceso de

extracción de información de una red social. Para ello no es necesario pertenecer a la red social, puesto que accederemos de forma anónima.

Para realizar la extracción se parte de una base de datos de la entidad financiera sobre un Excel, teniendo en la columna A el nombre, en la columna B los apellidos, y en la columna C la ubicación que tiene registrada el banco para cada cliente. El programa seguirá entonces el siguiente proceso:

Step 1: El programa lee el primer registro de la hoja de Excel, que contiene datos referentes a nombre, apellidos y ubicación.
Step 2: Se abre una página de Internet Explorer (IE) con la dirección web de la red social. Sobre esa página, el programa se situará sobre los campos de búsqueda. En el campo reservado para el nombre, escribirá el nombre, haciendo lo mismo con los apellidos. Después presionará el botón de búsqueda.
Step 3a: En caso de que sólo haya un perfil que coincida con el nombre y los apellidos buscados, se mostrará directamente dicho perfil. Entonces el programa compara la ubicación contenida en el perfil con la ubicación leída en la base de datos.
Step 4a: Si la ubicación coincide, descargará el perfil, cerrará el navegador y reiniciará el proceso con el siguiente registro. En caso contrario, se reiniciaría el proceso con el siguiente registro sin descargar la información.
Step 3b: En caso de haber más de un perfil que coincida con el nombre y apellidos buscados, la red social mostrará una lista de resultados. Entonces el programa recorrerá la lista de resultados, comparando la ubicación de cada perfil con la ubicación que ha leído de la base de datos.
Step 4b: Cuando el programa encuentre la primera coincidencia, accederá al perfil, descargará la información solicitada, cerrará el navegador y reiniciará el proceso con el siguiente registro. En caso de no encontrar ninguna coincidencia, no accederá a ningún perfil. Cerrará el navegador y reiniciará el proceso para el siguiente perfil.
Step 5: Una vez se hayan agotado los registros, fin del programa.

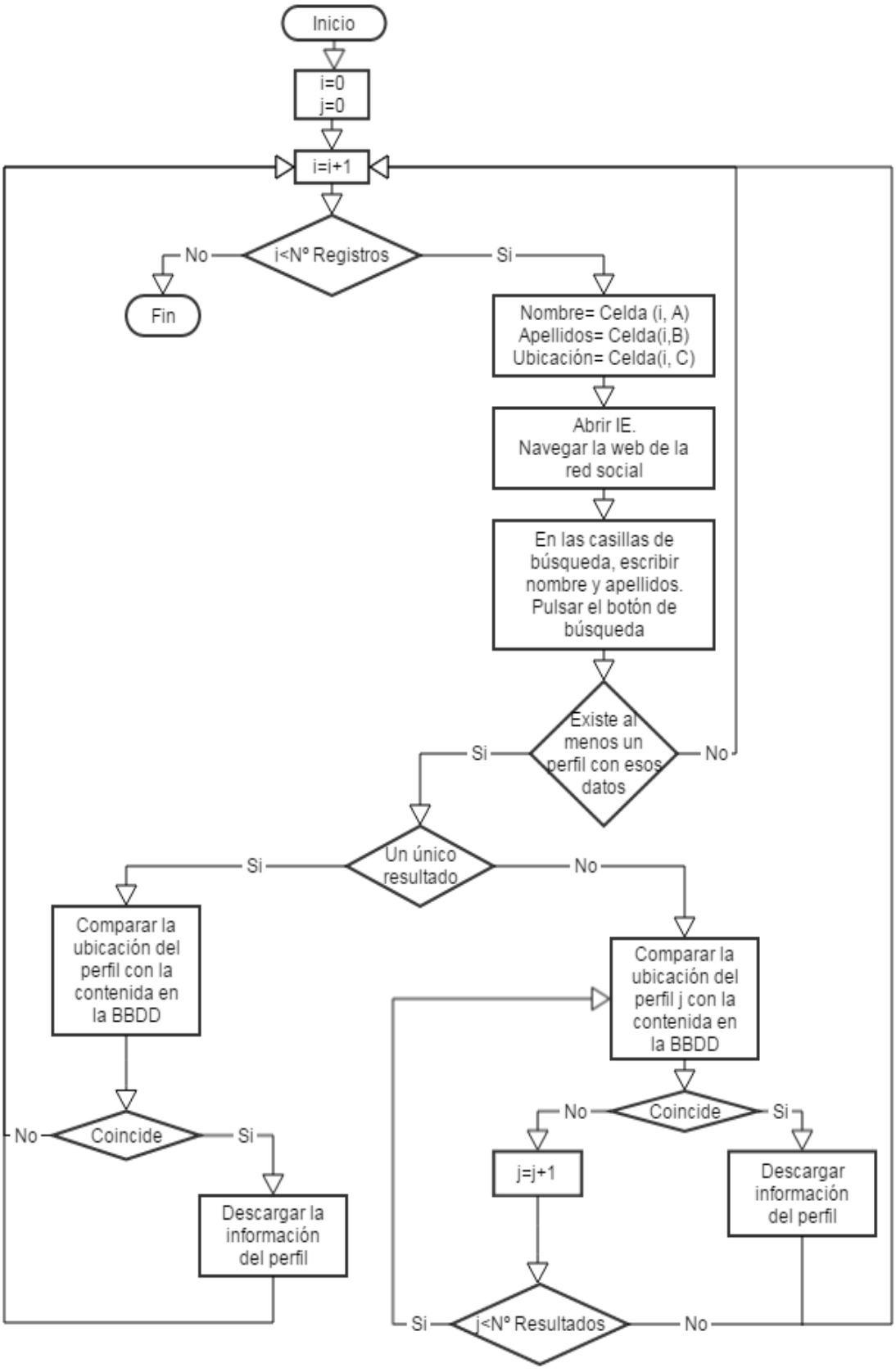

Figura 3Diagrama de flujo del programa desarrollado

3.2.2 Consideraciones y mejoras sobre la extracción

En esta sección se detallarán algunas mejoras aplicables a la extracción de datos. Sobre el método empleado, la herramienta VBA no es del todo estable. Es posible que en ocasiones se detenga sin motivo aparente, o haya veces que el código dé error en mitad de un proceso. Con un programa más potente tendríamos menos pausas y menos problemas durante la extracción. Como se ha detallado anteriormente, al buscar nombre y apellido comparamos los resultados con la ubicación, con lo que se abre un abanico de posibles errores:

- Si hay dos personas con mismos nombre y apellidos, y las dos residen en la misma ubicación, se estará extrayendo la que primero aparezca en LinkedIn, el cual no tiene por qué ser el perfil que se busca.
- Supongamos que hay al menos dos personas con mismos nombre y apellidos que residen en la misma zona. Si una tiene LinkedIn y otra no, se extraerá la que tiene perfil en LinkedIn cuando se puede estar buscando a la que no lo tiene.
- Al acceder a LinkedIn sin estar *loggeados*, es decir, de manera anónima, la propia red social limita las búsquedas en función del nivel de privacidad de cada usuario. Por esta razón se estará obteniendo menos información de la que realmente hay en internet, pero es información inaccesible con éste método. (Dentro de los errores posibles, éste sería el más leve, ya que se estaría descargando menor cantidad de información, pero no asignando información incorrecta).
- Al acceder a LinkedIn de manera anónima, no se muestra toda la información de un perfil, sino únicamente la información pública. La descarga del resto de la información de manera anónima no es posible.

Dadas las limitaciones temporales y económicas, y a pesar de ser conscientes de las limitaciones de la extracción, se ha decidido continuar el proyecto con el método anteriormente descrito, para, en caso de resultado satisfactorio, mejorar la calidad de la extracción.

3.3 Base de datos de riesgos de una entidad bancaria

Para la realización del estudio se partirá de una base de datos de riesgos cedida por una entidad financiera. Dicha base de datos contiene la información de un modelo experto de Scoring, con el cual se evalúa cada operación asociada a cada cliente. Por lo tanto, salidas reales del modelo, se dispone sólo en el caso en el que el cliente haya solicitado un producto financiero determinado.

En el momento de iniciar este trabajo, se esperaba que la base de datos contuviera un campo con los datos de las nóminas que los clientes tienen domiciliadas en la entidad. Con dicho campo, se podría realizar una descarga más precisa de la red social, pero las bases de datos de las entidades financieras son altamente confidenciales, por lo que dicho campo viene blanqueado. Se dispone de nombre y apellidos de cada cliente, localización, y una calificación de morosidad asociada a cada cliente. Dicha clasificación puede verse en la Tabla 2.

	Categorización del cliente según riesgo de crédito
1	Riesgo normal
2	Riesgo Subestándar
3	Riesgo Dudoso / en default por razón de morosidad / retraso superior a 90 días del cliente
4	Riesgo Dudoso / en default por razones distintas de morosidad / retraso del cliente
5	Riesgo Fallido / Castigado

Tabla 2 Categorización del cliente según el riesgo de crédito

Sobre dicha categorización, se creará una variable dicotómica en nuestra base de datos, la cual valdrá 0 si el código es 1 o 2 (Cliente sano), y valdrá 1 si el código es 3, 4 o 5 (Cliente moroso).

La base de datos de partida contiene aproximadamente 27.000 registros, cada uno con información referente a un cliente del banco. Una vez buscados en la red social, resulta una base de datos de aproximadamente 2.600 registros. Sobre esta última base de datos se comienza el ejercicio.

4. Tratamiento de datos

4.1 Introducción: Data Science

Todos los días se generan cantidades colosales de datos. Cada vez es más rápida su obtención y cada vez se almacenan cantidades mayores porque es más fácil recolectarlos y más barato guardarlos. Este fenómeno de almacenamiento masivo de información es conocido como Big Data, y algunas evidencias al respecto son (Marr, 2015):

- Más del 90% de todos los datos que hoy existen han sido creados en los dos últimos años.
- La capacidad de procesamiento se ha multiplicado por 300 desde el año 2000, permitiendo procesar millones de transacciones por minuto.
- Se espera que en 2015 se creen 1.9 millones de puestos de trabajo en EEUU relacionados con las tecnologías de la información IT. Se estima que cada uno de esos puestos genere a su vez 3 puestos de trabajo no relacionados con las IT, y todo gracias al Big Data.

Dicho término engloba además todas las técnicas relacionadas con los flujos masivos de datos, como la Minería de Datos o la Ciencia de los Datos (*Data Science*).

La obtención de dichos datos es comúnmente llamado *Data Mining* o Minería de Datos. Automatiza procesos para identificar patrones o tendencias, logrando descubrir relaciones que de otro modo serían imposibles de deducir dada la gran cantidad de datos. Las obras más representativas sobre este tema son (Hernández-Orallo, Ferri, Lachinche, & Flach, 2004) y (Aluja, 2001). Se trata por tanto de la metodología empleada para obtener y tratar gran cantidad de información, utilizando técnicas de estadística, inteligencia artificial, aprendizaje automático y sistemas de bases de datos.

Una vez se genera tal flujo de información, se denomina *Data Science* a la extracción de conocimiento de dichos datos. Supone por tanto un avance sobre las técnicas de *Big Data*, ya que además de analizar los datos, es posible interpretar las relaciones entre los datos, generando valor. Además de las técnicas de *Big Data*, un *Data Scientist* (literalmente, Científico de Datos) debe tener un amplio conocimiento del negocio, que será clave en la conducción de la investigación. En el caso de los modelos en el sector financiero, será decisivo el conocimiento de negocio para la correcta implantación del modelo en la gestión de un banco. Los equipos de *Data Science* son equipos multidisciplinares que suelen estar compuestos por matemáticos, ingenieros, informáticos, economistas o científicos.

Se trata de una actividad en expansión ya que casi cualquier sector empresarial se puede aprovechar de su utilización. Sus aplicaciones se extienden a campos como la medicina, las compañías aéreas, el sector financiero, aseguradoras o empresas de venta al por menor.

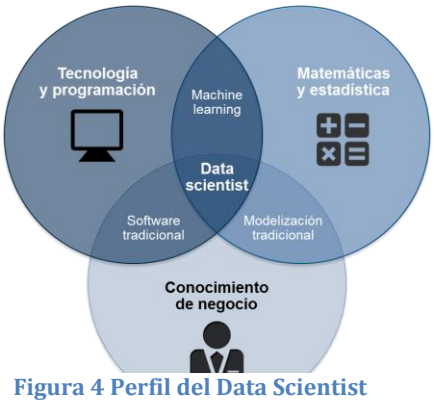

Figura 4 Perfil del Data Scientist

Así, gracias al *Big Data*, las empresas pueden ofrecer servicios más personalizados, evitar pérdidas o fugas de clientes, optimizar la producción o decidir sobre diferentes alternativas.

Un proceso de Data Science suele seguir los pasos que se ilustran en la figura 5.

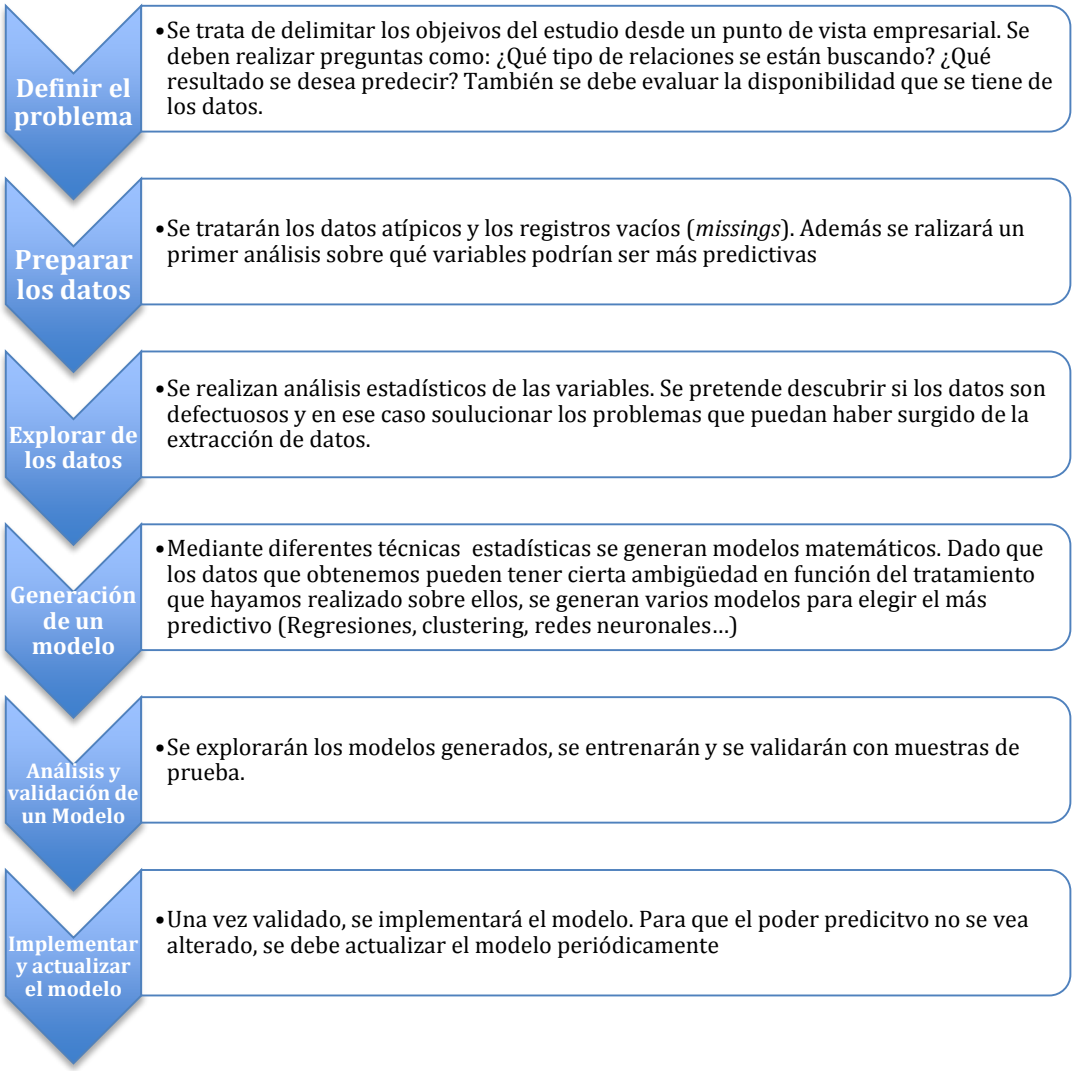

Figura 5 Diagrama de flujo de un proceso de Data Science

4.2 Tratamiento de los datos

Los datos descargados de la red social no vienen normalizados. Cada registro de la base de datos contiene información escrita de distinta forma y a menudo con faltas de ortografía. Para poder tratar toda la información será volcaremos en SAS, pero previamente debe ser normalizada y categorizada, pues sino al automatizar el proceso no se obtendrán los resultados deseados. También nos apoyaremos en Excel en algunos casos.

En esta sección se explicará el tratamiento realizado a los datos para poder procesarlos con SAS. Se describirá el proceso de normalizado de cada campo extraído de la red social por separado ya que cada campo se normalizará de forma diferente.

4.2.1 Normalizado del campo 'U*niversidades*'

A la hora de normalizar los datos de las universidades, se debe agregar bajo un mismo título cada universidad. Para el programa, '*Universidad Complutense de Madrid*', '*Universidad Complutense de Madrid (UCM)'* o '*UCM'* no son lo mismo. Se debe reescribirlas todas bajo el título que se prefiera.

Para el tratamiento de datos existen dos opciones: o bien escribir un código en SAS especificando que los campos que *suenen como 'UCM'* o *suenen como 'Complutense Madrid'* los categorice en '*Universidad Complutense de Madrid*', o bien tratar los datos con Excel. Dada la extensión de los datos y la variedad en cuanto a universidades, hemos optado por Excel. Mediante filtros, se seleccionan las universidades similares y se reescriben directamente sobre Excel. Una vez terminado el proceso, se importará esa base de datos en SAS.

4.2.2 Normalizado del campo 'E*mpresas'*

Para normalizar las empresas, se categorizarán por su ámbito de actuación en:

- Empresas locales: Operan dentro de una localidad determinada.
- Empresas nacionales: Operan por parte o todo el territorio Español.
- Empresas internacionales: Operan en varios países, incluyendo España

Para clasificar cada empresa en una categoría de las anteriores, se escribirá un código en SAS. El código que procesa toda la información correspondiente a nombres de empresa es sencillo pero muy extenso. Se crea un bucle que buscará para cada registro palabras que *suenen como* una sentencia dada. Debido a la gran variedad de empresas y maneras en las que los usuarios lo escriben en las redes sociales, se genera una casuística vastísima que ha de ser tratada. Por ejemplo, para el programa no es lo mismo *L'ORÉAL* que *LORÉAL* que *L'OREAL* o *LOREAL*. Si extendemos esto a una base de datos con más de 7.000 registros, prácticamente se obtienen 7.000 empresas diferentes. No tendría sentido detallar todo el código en estas líneas, éstos serían los ejemplos más básicos:

- Campos que contengan palabras como '*ayuntamiento'* nombres de provincia o campos con palabras como '*e hijos'* se agrupan masivamente en *LOCAL*.
- Se agrupan campos con palabras como '*nacional*', '*ministerio'* o '*consejo'* como *NACIONAL.*
- Se agrupan campos que contengan palabras como '*group', 'international', 'internacional', 'multinacional'* en *INTERNACIONAL.*

No obstante, se debe recorrer la tabla buscando casos diferentes e intentar buscar un nivel de agregación que incluya a la mayoría.

4.2.3 Normalizado del campo 'C*argos'*

Se ha tratado todo el campo *Cargos* agregando lo máximo posible para poder tener una variable que pueda ser estudiada. Al igual que en el caso de las empresas, resulta un proceso lento tratar este tipo de datos, ya que son muchas las formas de escribir una misma posición dentro de una empresa (P.E: el cargo *senior* no tiene la misma categoría en todas las empresas). Sumado a todo esto, abundan las faltas de ortografía que terminan de complicar el tratamiento de datos.

4.2.4 Normalizado de los campos *'Fechas de estudios' y 'Fechas de trabajos'*

Para poder tratar las fechas también deben ser normalizadas. Para ello se utilizará Excel, y separando cada fecha en una columna diferente, las se clasifican como fecha de inicio o fecha de fin. Con esta división de fechas será más rápido el análisis de variables como por ejemplo *Tiempo máximo en un empleo* o *Tiempo máximo en una empresa*.

4.2.5 Normalizado del campo *'Idiomas'*

Los idiomas se normalizarán reescribiendo el campo, de forma de '*Español*', '*Castellano*' o '*Spanish*' sean todos '*ESPAÑOL*'. Se debe realizar un análisis de los datos, buscando todas las formas posibles de escribir cada idioma para agruparlo bajo una única categoría. Para ello se ha desarrollado un código en SAS similar a los códigos antes descritos: Realiza una búsqueda por el campo '*Idiomas*', y todo aquello que suene como '*Inglés*',' *Ingles*', '*English*', '*Inglish*' será reescrito como '*INGLES*' (como también se comentó anteriormente, abundan las faltas de ortografía en las redes sociales)

4.2.6 Categorización del campo *'Sector'*

El campo *'Sector'* será el único que ya viene normalizado en la base de datos. Sin embargo será necesario categorizarlo, ya que suele ser un campo con alto nivel de predictibilidad. Para ello, se categoriza de la siguiente forma:

- Por un lado se agregan los 140 sectores iniciales en 19, renombrados con letras de la 'A' a la 'S'. De esta forma tendremos los datos más agregados y podremos trabajar con ellos (Tabla 3)
- Por otro lado, se agregan dependiendo del salario medio del sector (Tabla 3), agrupados en cuatro categorías (INE, 2013):
    1. Salario medio<1.500€
    2. Salario medio comprendido entre 1.500€ y 2.000€
    3. Salario medio comprendido entre 2.000€ y 2.500€
    4. Salario medio>2.500€

Salarios medios mensuales brutos por actividad.

Unidades: euros.

	2011	2012	2013
Total	1.841,8	1.850,3	1.869,1
A Agricultura, ganadería, silvicultura y pesca	1.211,9	1.160,0	1.175,0
B Industrias extractivas	2.765,5	2.528,6	2.224,3
C Industria manufacturera	2.112,8	2.147,7	2.192,7
D Suministro de energía eléctrica, gas, vapor y aire acondicionado	3.357,1	3.497,5	3.391,1
E Suministro de agua, act. de saneamiento, gestión de residuos y descontaminación	2.118,4	2.063,1	2.098,7
F Construcción	1.748,8	1.823,0	1.897,1
G Comercio al por mayor y al por menor; reparación de vehículos de motor y motocicletas	1.489,6	1.497,1	1.497,5
H Transporte y almacenamiento	1.863,3	1.863,6	1.844,5
I Hostelería	1.222,3	1.200,5	1.180,6
J Información y comunicaciones	2.445,5	2.476,8	2.602,8
K Actividades financieras y de seguros	2.004,0	1.708,1	1.645,9
L Actividades inmobiliarias	2.004,0	1.708,1	1.645,9
M Actividades profesionales, científicas y técnicas	1.972,8	2.096,2	2.039,5
N Actividades administrativas y servicios auxiliares	1.260,3	1.278,8	1.292,5
O Administración Pública y defensa; Seguridad social obligatoria	2.304,0	2.286,3	2.420,1
P Educación	2.328,6	2.219,6	2.256,9
Q Actividades sanitarias y de servicios sociales	2.113,6	2.129,9	2.115,2
R Actividades artísticas, recreativas y de entretenimiento	1.501,5	1.457,2	1.470,9
S Otros servicios	1.320,2	1.289,2	1.280,6
T Act. De los hogares como empl. de pers doméstico y prod. de bienes y serv. para uso propio	773,4	781,8	709,0

Tabla 3 Salarios medios brutos por nivel de actividad. Fuente: (INE, 2013)

4.1.7 Mejoras sobre la normalización

En primer lugar, se debe entender que al extraer los datos de una red social, cada usuario puede escribir en su perfil lo que considere, lo que da lugar a un abanico inmenso de posibilidades sobre el mismo campo. Si se dispusiera de más tiempo, más presupuesto y mejores medios, se podría normalizar la base de datos con métodos más automáticos, pero para el presente estudio se ha considerado que la mejor opción es el normalizado de datos mediante códigos que reescriben cada campo. Con todo y con ello, se podrían realizar una serie de mejoras:

- Sobre el campo '*Cargos*': Se podría categorizar más exhaustivamente los cargos, con un programa más detallado.
- Sobre el campo '*Empresas*': lo ideal sería disponer de una base de datos con el mayor número de empresas posible, para poder buscar el dato que se ha extraído en esa base de datos y tener así información sobre el tamaño, ámbito de operación, o sector de actividad. Estas bases de datos existen, están en venta, pero dado el presupuesto del proyecto no ha sido posible disponer de ellas.

- Sobre el campo 'U*niversidades'*: Se podría disponer de los estudios de cada perfil, categorizarlos por ramas y obtener más información.
- Sobre el campo '*Idiomas':* Con técnicas de Procesamiento de Lenguaje Natural, el programa podría entender que el nombre de un idioma, escrito en diferentes idiomas se refieren a lo mismo, y por lo tanto constituyen el mismo valor para una variable.

Además, sería altamente interesante poder tratar toda la base de datos con el mismo programa, presumiblemente SAS. La opción más automática sería anexionar todos los códigos antes descritos para que el programa corriera una vez, y se obtuviera como salida la base de datos normalizada.

4.3 Creación de variables

Se realiza un análisis experto para decidir qué variables se espera que sean predictivas, atendiendo al sentido de negocio. Para el estudio inicial, se plantean las variables que podrían tener relación con la morosidad. En fases posteriores del estudio se analizará cuantitativamente el impacto que estas variables tienen en el modelo. Así, las variables que se crearán vienen detallas en la Tabla 4.

Además, se tratarán los *missings* de la siguiente forma: para los campos vacíos en las variables categóricas, se dará el valor de la moda; para las variables continuas, se imputarán con la mediana.

	Idiomas:
1	*Indicador '¿Habla idiomas (sin contar español ni otros idiomas cooficiales del Estado)?' (1='Sí';0='No')*
2	*Indicador '¿Habla idiomas (sin contar español y contando otros idiomas cooficiales del Estado)?' (1='Sí';0='No')*
3	*Número de idiomas (asumiendo que español es uno y sin contar con otros idiomas cooficiales del Estado)*
4	*Número de idiomas (asumiendo que español es uno y contando con otros idiomas cooficiales del Estado)*
5	*Indicador '¿Habla inglés?' (1='Sí';0='No')*
6	*Indicador '¿Habla francés?' (1='Sí';0='No')*
	Sector:
1	*Sector de empleo - Categorizada en función de la tipología 1*
2	*Sector de empleo - Categorizada en función de la tipología 2*
3	*Sector de empleo - Categorizada en función de una aproximación del nivel salarial*
	Universidad:
1	*Indicador '¿Tiene estudios?' (1='Sí';0='No (Sólo desinformados)')*
1 bis	*Indicador '¿Tiene estudios?' (1='Sí';0='No (Desinformados + Educación secundaria)')*
2	*Indicador '¿Tiene estudios superiores?' (1='Sí';0='No')*
3	*Duración estudios (total)*
4	*Tiempo desde la última vez que estudió*
	Cargo/Empresa:
1	*Indicador '¿Tiene trabajo (actualmente)?' (1='Sí - asalariado';0='No - no asalariado (ni pensionistas ni parados)')*
1 bis	*Indicador '¿Tiene trabajo (actualmente)?' (3='Trabajador';2='Parado';1='Estudiante';0='Jubilado')*
1 ter	*Indicador '¿Tiene trabajo (actualmente)?' (1='Estudiantes + Parados';0='Trabajadores y jubilados')*
2	*Número de cargos/trabajos históricamente*
3	*Antigüedad laboral (periodo de tiempo desde el primer cargo)*
4	*Duración del último cargo*
5	*Duración máxima en un cargo*
6	*Duración mínima en un cargo*
7	*Duración media en sus cargos*
8	*Indicador '¿Ha estado (en algún momento) sin empleo (dos o más meses, por margen de error)?' (1='Sí';0='No')*
8 bis	*Indicador '¿Ha estado (en algún momento) sin empleo durante más de seis meses?' (1='Sí';0='No')*
8 ter	*Indicador '¿Ha estado (en algún momento) sin empleo durante más de un año?' (1='Sí';0='No')*
9	*Máximo periodo de tiempo sin cargos/trabajo*
10	*Máximo cargo asumido (Directivo o gerente vs. técnicos o empleados)*
11	*Duración del máximo cargo asumido*
12	*Tipo de empresa (Local/Nacional/Internacional)*
	Datos Personales:
1	*Localidad*

Tabla 4 Variables creadas para el estudio

Aunque la creación de dichas variables puede parecer trivial, requieren un nuevo tratamiento de los datos.

Los indicadores serán las variables más sencillas de construir. Serán variables artificiales o *dummies,* que devolverán el valor 1 si el valor de la variable coincide con un valor dado, y 0 en caso contrario. Por ejemplo, si el campo '*Idiomas'* contiene '*FRANCES'*, el indicador '*¿Habla francés?'* devolverá un 1. Para las variables como '*Número de idiomas'* se contarán los registros de idiomas que tiene cada cliente.

Sin embargo, las variables del estilo '*duración máxima/mínima'* requieren un tratamiento más completo: se debe sacar la duración de cada cargo (fecha de fin – fecha de inicio) y escoger el máximo o el mínimo. Para la variable '*máximo periodo de tiempo sin trabajo'* se resta la fecha de inicio del cargo *i* menos la fecha de comienzo del cargo *i-1*. Para la variable '*Duración media en sus cargos'* se halla la duración en cada uno de sus cargos y se calcula su media.

Con las variables antes descritas, se construye una única tabla que contendrá toda la información sobre la cual se creará el modelo. Esta tabla contiene 43 campos (columnas de las tablas) donde 39 son variables de estudio. La tabla se compone de 2.626 registros. Sobre dicha tabla se efectuarán los análisis (sección 6).

5 Modelo de Regresión Logística

En esta sección se define el modelo de regresión logística teórica, para dar paso a la construcción del modelo en la sección 6.

El extenso uso de modelos de regresión logística en el sector financiero, nace de la necesidad de evaluar el riesgo (como se ha detallado en secciones anteriores) y la naturaleza del modelo: se obtiene una respuesta binaria y resulta un modelo muy fácil de interpretar. Es por ello que los modelos de regresión se han convertido en una herramienta básica en cualquier análisis de datos que involucre la descripción de la relación entre una variable de respuesta y una o más variables explicativas (Llaugel & Fernández, 2001).

5.1 Formulación del modelo

Un modelo de regresión logística o Logit es un modelo de respuesta binaria, en lo sucesivo Y. La variable Y sigue una distribución de Bernouilli de parámetro p ∈ [0, 1]. Se desea construir un modelo Y(x) en función de los parámetros y un error. El modelo de regresión logística es de la forma:

$$p(x_i) = \pi_i = \frac{e^{(\alpha+\sum_j B_j x_{ij})}}{1 + e^{(\alpha+\sum_j B_j x_{ij})}}$$

Tomando el logaritmo natural, se obtiene:

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = a + \sum_j b_j x_{ij}, \qquad (i = 1, \dots, N), (j = 1, \dots, K)$$

Donde K es el número de variables independientes, N el número de observaciones en la muestra, $x_i = (x_{i1}, \dots, x_{iK})$ el vector que contiene las observaciones de cada variable para el individuo i-ésimo, y $B = (b_1, b_2, \dots, b_K)$ el vector que contiene los parámetros del modelo. El cociente $\frac{\pi_i}{1-\pi_i}$ representa la ventaja de respuesta Y=1 para los valores observados de las variables independientes.

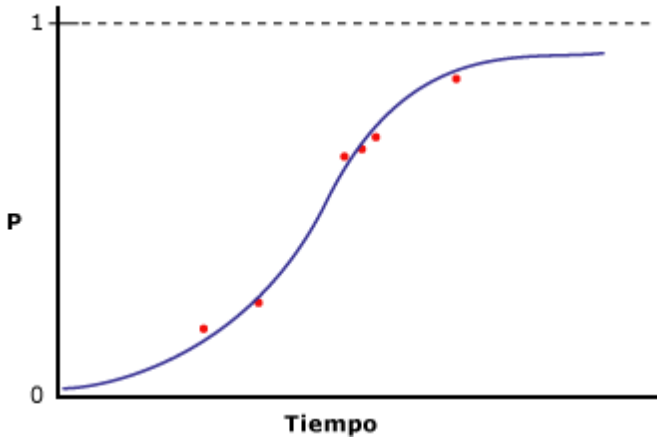

Figura 6 Ejemplo de modelo Logit

El modelo Logit se encuentra acotado en el eje de coordenadas en el intervalo [0,1] y es asintótico en y=0 e y=1. En este modelo, todos los parámetros b_i son lineales, y dependiendo del rango de x, pueden ir desde - ∞ hasta +∞ (Hosmer, Lemeshow, & Sturdivant, 2013)

5.2 Estimación

El modelo de regresión logística se estima a través del método de máxima verosimilitud (MV), que asigna máxima probabilidad a los datos observados. La función de verosimilitud de los datos respecto a los parámetros del modelo logit, que ha de ser maximizada, es de la forma:

$$\prod_{q=1}^{Q} \binom{n_q}{y_q} p_q^{y_q} (1 - p_q)^{n_q - y_q}$$

Siendo $x_q = (x_{q0}, x_{q1}, \dots, x_{qK})$ con $q = (1, \dots, Q)$ la q-ésima combinación de valores de las K variables explicativas de la muestra de tamaño N. Calculando el logaritmo del núcleo de la función anterior, derivando respecto a cada parámetro b_i e igualando a 0, se obtiene la ecuación de verosimilitud:

$$\sum_{q=1}^{Q} y_q x_{qk} - \sum_{q=1}^{Q} n_q \hat{p}_q x_{qk} = 0, \quad k = 0, \dots, K$$

Con el estimador MV de p_q,$\hat{p}_q$ y los estimadores de MV de b_i, $\hat{b}_i$:

$$\hat{p}_q = \frac{e^{\sum_{r=0}^{R} \hat{b}_r x_{qr}}}{1 + e^{\sum_{r=0}^{R} \hat{b}_r x_{qr}}}$$

Para resolver las ecuaciones de máxima verosimilitud se hace uso del método iterativo de Newton-Raphson, con el cual se busca un cero o un máximo de una función.

En dicho método, se parte de un valor de x cercano a 0, y se calcula la tangente de la función en ese punto. El punto de corte de dicha tangente con el eje de abscisas será un valor más cercano a la raíz de la función. Se itera hasta que el punto de corte de la tangente coincida con el punto en el cual se evalúa la función. De este modo, el método se expresa como: $x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$.

5.3 Inferencia sobre los parámetros y bondad de ajuste del modelo

En este trabajo se realizará el contraste condicional de razón de verosimilitudes con los correspondientes intervalos de confianza para cada b_i.

En el contraste condicional de razón de verosimilitudes se parte de un modelo M_1 que se ajusta bien, y se desea contrastar si un subconjunto de los parámetros b_i, denotado $C = (C_1, \dots, C_l)'$, es nulo. Al hacer 0 esos parámetros en M_1 se obtiene un modelo M_2 anidado a M_1. La hipótesis nula es C=0 frente al modelo completo M y el estadístico de contraste verifica que: $G^2(M_2|M_1)=G^2(M_2)-G^2(M_1)$, es decir, es la diferencia de los contrastes de razón de verosimilitudes de bondad de ajuste para cada modelo. Éste tiene, bajo el modelo M_2, distribución χ^2 con l grados de

libertad. Se rechazará la hipótesis con nivel de significación α si el valor estadístico es mayor o igual que el cuantil 1-α de la distribución.

En resumen, se realizará un contraste para cada parámetro, en los cuales se evalúa si cada parámetro b_i es significativamente distinto de 0. Además dado que b_i sigue una N (b_i, $\hat{\sigma}(\hat{b}_i)$), se construyen intervalos de confianza a nivel 1-α para cada b_i, que resultan:

$$b_i \pm z_{\frac{\alpha}{2}} \hat{\sigma}(\hat{b}_i).$$

Para medir la bondad de ajuste del modelo haremos uso del área bajo la curva ROC y la Tasa de Clasificaciones Correctas.

La Tasa de Clasificaciones Correctas se calcula como el cociente entre los aciertos y el tamaño muestral. Se debe establecer un punto de corte, que aunque normalmente sea 0.5, en ocasiones es más apropiado tomar la proporción de valores Y=1 de la muestra. Otra opción es tomar varios puntos de corte, y escoger el que mayor porcentaje de aciertos obtenga.

La curva ROC se construye representando la tasa de verdaderos positivos frente a falsos positivos. Por lo tanto, cuando mayor sea esa medida, mejor será el modelo, pudiendo considerarse el modelo preciso cuando el área bajo la curva supere 0.7.

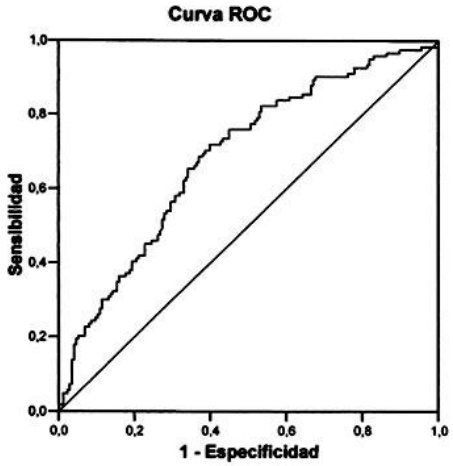

Figura 7 Ejemplo de Curva ROC

5.4 Selección de variables

Para la correcta interpretación del modelo, se debe escribir la regresión con el mínimo de variables posibles, atendiendo además al sentido de negocio que generen dichas variables.

Existen tres métodos de selección de variables (Siddiqqi, 2006):

- *Forward Selection:* Se comienza con un modelo que contenga únicamente la variable más predictiva y se itera añadiendo variables mientras no perjudique el nivel de predictibilidad
- *Backward Elimination:* Se introducen todas las variables en el modelo, y se itera eliminando variables mientras aumente el nivel de predictibilidad.
- *Stepwise:* Se ajusta un modelo para cada una de las variables, y se selecciona para entrar en el modelo aquella variable que origine el mejor modelo. A continuación, se ajusta un modelo por cada una de las variables restantes, añadiendo la variable seleccionada. Se itera este proceso mientras aumente el poder predictivo. Difiere del *Step Forward* en que con este método se pueden eliminar variables en algún paso.

En este trabajo se seleccionarán variables mediante el método *Stepwise*. En los pasos en los que se añaden nuevas variables, se realizan contrastes de razón de verosimilitudes con el modelo del paso anterior como hipótesis nula y cada uno de los nuevos como hipótesis alternativa en cada contraste. Aquellos contrastes cuyo P-Valor sea inferior al nivel de significación requerido determinan las variables susceptibles de ser introducidas en el modelo en ese paso (Couso, 2011).

5.5 Diagnosis del modelo: análisis de los residuos

Una vez ajustado el modelo de regresión, un paso fundamental es la validación del mismo, mediante un análisis exhaustivo de sus residuos. Trabajaremos con los residuos de la devianza, y en caso de que alguno sea significativo se analizará su influencia mediante las distancias de Cook. Los residuos de la devianza están basados en el estadístico G^2 y se definen como:

$$d_q = signo(y_q - \hat{m}_q)(2\left[y_q \ln\left(\frac{y_q}{\hat{m}_q}\right) + (n_q - y_q)\ln\left(\frac{n_q - y_q}{n_q - \hat{m}_q}\right)\right])^{1/2}.$$

Estos residuos pueden ajustarse a una N (0, 1). Lo que haremos será contrastar si los residuos son cero (H_0) frente a que sean significativamente distintos de cero (H_1). Para ello, para aquellos residuos que en valor absoluto exceden de 2 rechazamos H_0.

La distancia de Cook mide la influencia de una observación sobre la estimación de los parámetros del modelo. Si no hay ninguna medida mayor que 1, y existen residuos que en valor absoluto exceden de 2, se podrán aceptar dichos residuos y dar por validado el modelo. En cambio, si la distancia de Cook es mayor que uno, se debe eliminar el registro que contiene el residuo problemático y reajustar el modelo.

La distancia de Cook se define como:

$$COOK_i = \frac{1}{p}(\hat{\beta} - \hat{\beta}_i)'X'WX(\hat{\beta} - \hat{\beta}_i).$$

5.6 Interpretación

El modelo de regresión logística se interpreta según los coeficientes b_i junto a la constante "a".

La constante es el valor del logaritmo de la ventaja de respuesta Y=1 para un individuo que tiene valor 0 en todas las variables predictivas, o bien cuando la respuesta no depende de ninguno de los parámetros de predicción.

A partir de los parámetros b_i se pueden calcular los "cocientes de ventajas". Suponiendo un modelo en el que únicamente se obtuviera un coeficiente b_i distinto de 0, el riesgo relativo de respuesta Y=1 para dos valores distintos x_1y x_{12} del predictor, se define como $R_{12} = \frac{p(x_1)}{p(x_2)}$ y el cociente de ventajas de respuesta Y=1 dados los mimos valores, se define como:

$$\theta_{12} = \frac{\frac{p(x_1)}{1-p(x_1)}}{\frac{p(x_2)}{1-p(x_2)}}$$

Ambos conceptos se relacionan según:

$$\theta_{12} = R_{12} * \frac{1-p(x_2)}{1-p(x_1)}$$

De esta forma, el riesgo relativo puede ser aproximado mediante el cociente de ventajas si la probabilidad de Y=1 es cercana a 0. Se interpretarán, por tanto, los parámetros del modelo en función del cociente de ventajas. Dichos cocientes de ventaja se denominan *Odds Ratio* (OR).

Si la variable de estudio es continua, el cociente de ventajas representa la variación en la ventaja de respuesta Y=1 por cada unidad de aumento de la variable cuando el resto permanecen constantes. La ventaja de respuesta Y=1 queda multiplicada por el exponencial de b_i. Por lo tanto:

- OR=1 significa que dicha variable en concreto no afecta a la respuesta.
- OR<1, la ventaja de respuesta disminuye con el aumento de b_i
- OR>1, la ventaja de respuesta Y=1 aumenta al aumentar la variable.

Si la variable es categórica, se tiene una OR que representa la ventaja de respuesta de esa categoría en concreto con respecto a la categoría de referencia, cuando el resto de las variables se quedan con valor fijo.

6 Análisis del modelo de predicción de impagos

6.1 Análisis descriptivo univariante

En esta sección se lleva a cabo el estudio estadístico de cada una de las variables candidatas con el fin de determinar incidencias, tratar valores atípicos, o reagrupar variables categóricas. Tras un primer análisis de las variables, se estudiarán por separado las variables continuas de las categóricas.

6.1.1 Análisis de variables categóricas

Para el análisis de las variables categóricas se ha empleado el programa Statgraphics. Con el procedimiento *"Cross Tabulation"* obtenemos un gráfico en el que se muestra la distribución de cada variable según el indicador de morosidad.

A continuación se detalla el análisis de las variables consideradas para la construcción del modelo.

En general, la tasa de morosos en nuestra muestra es inferior a la de clientes sanos, como es lógico en una entidad financiera. Se tienen 418 clientes morosos en una muestra de 2.626 clientes.

En el estudio de *Número de trabajos históricos* según el indicador de mora, observamos una distribución similar en los dos casos. Sin embargo, como se muestra en la tabla 6, la probabilidad condicionada de que, en la muestra, alguien con más de 6 trabajos sea moroso, es sustancialmente superior a las otras categorías.

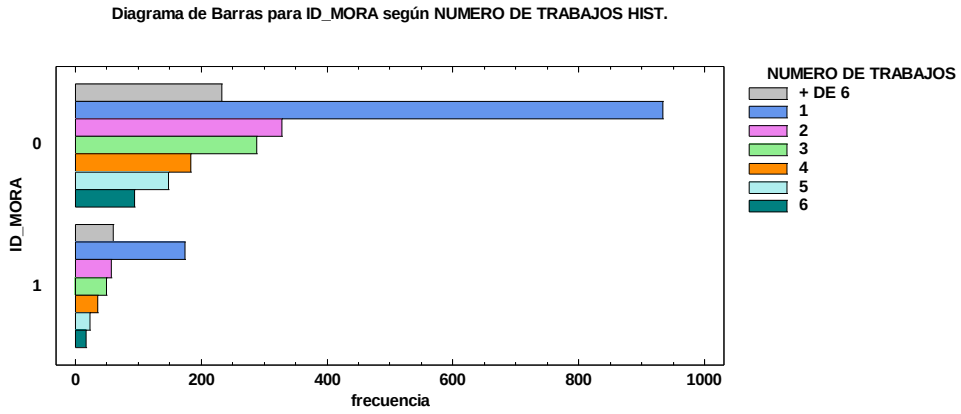

Figura 8 Número de trabajos históricos según Indicador de mora

	+ DE 6	1	2	3	4	5	6	Total por Fila
0	233	933	328	288	183	148	95	2208
	8,87%	35,53%	12,49%	10,97%	6,97%	5,64%	3,62%	84,08%
1	60	174	58	50	36	23	17	418
	2,28%	6,63%	2,21%	1,90%	1,37%	0,88%	0,65%	15,92%
Total por Columna	293	1107	386	338	219	171	112	2626
	11,16%	42,16%	14,70%	12,87%	8,34%	6,51%	4,27%	100,00%

Tabla 5 Resumen estadístico para Número de trabajos históricos según Indicador de mora

P(1|1)	P(1|2)	P(1|3)	P(1|4)	P(1|5)	P(1|6)	P(1|+ de 6)
0,1572	0,1503	0,1479	0,1644	0,1345	0,1518	0,2048

Tabla 6 Probabilidad condicionada de mora según el número de trabajos

En el análisis del sector de producción en el que opera la empresa donde trabaja el cliente según el indicador de mora, observamos de nuevo una distribución similar tanto para clientes sanos como para morosos. Estudiando la probabilidad condicionada de que, conociendo el sector de la empresa, un cliente de nuestra muestra sea moroso, se desprende que el sector primario tienen la menor probabilidad de default (11,76%) frente al sector cuaternario, con la mayor probabilidad de default, situada en 18,30%.

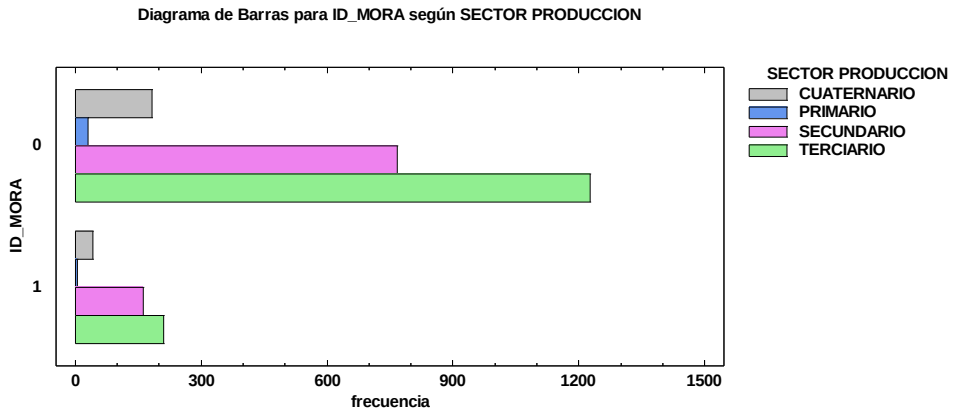

Figura 9 Sector de producción según Indicador de mora

	CUATERNARIO	PRIMARIO	SECUNDARIO	TERCIARIO	Total por Fila
0	183	30	767	1228	2208
	6,97%	1,14%	29,21%	46,76%	84,08%
1	41	4	163	210	418
	1,56%	0,15%	6,21%	8,00%	15,92%
Total por Columna	224	34	930	1438	2626
	8,53%	1,29%	35,42%	54,76%	100,00%

Tabla 7 Resumen estadístico para Sector de producción según Indicador de mora

P(1|Primario)	P(1|Secundario)	P(1|Terciario)	P(1|Cuternario)
0,1176	0,1753	0,1460	0,1830

Tabla 8 Probabilidad condicionada por el sector de producción

Para el estudio de la categoría salarial, se ha agrupado el salario medio del sector en el que trabaja el cliente en tres categorías:

S1: Sueldo medio del sector <1.500€
S2: 1.500<Sueldo medio del sector <2.500€
S3: Sueldo medio del sector >2.500€
Se observa que las distribuciones son similares.

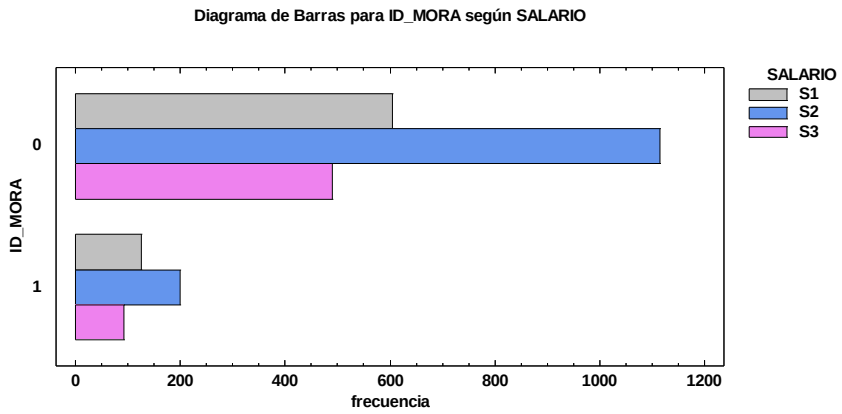

Figura 10 Salario según Indicador de mora

	S1	S2	S3	Total por Fila
0	604	1114	490	2208
	23,00%	42,42%	18,66%	84,08%
1	126	200	92	418
	4,80%	7,62%	3,50%	15,92%
Total por Columna	730	1314	582	2626
	27,80%	50,04%	22,16%	100,00%

Tabla 9 Resumen estadístico para Salario según Indicador de mora

En la Tabla 10 se puede ver que la probabilidad de Default es mayor a medida que el sueldo medio del cliente decrece.

P(1|S1)	P(1|S2)	P(1|S3)
0,1726	0,1522	0,1581

Tabla 10 Probabilidad condicionada de mora según el salario

Para el estudio del número de idiomas hablados por el cliente en función de la morosidad no encontramos una distribución diferente en las distintas categorías.

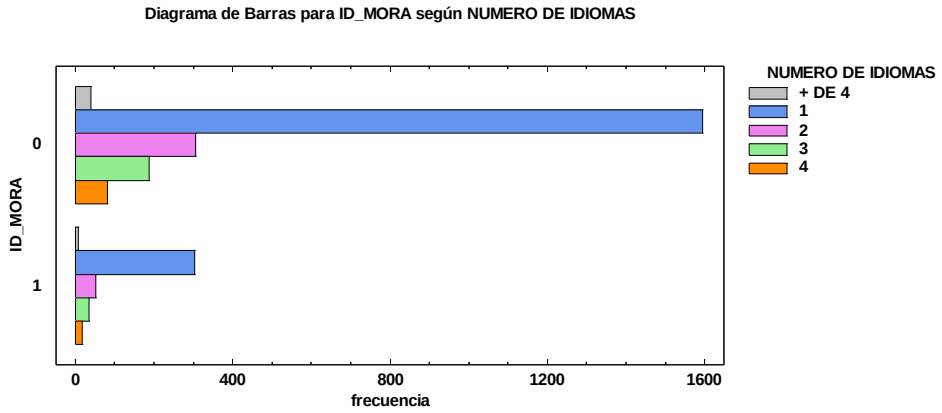

Figura 11 Número de idiomas según Indicador de mora

Tabla de Frecuencias para ID_MORA por NUMERO DE IDIOMAS

	+ DE 4	1	2	3	4	Total por Fila
0	40	1594	305	187	82	2208
	1,52%	60,70%	11,61%	7,12%	3,12%	84,08%
1	9	304	53	35	17	418
	0,34%	11,58%	2,02%	1,33%	0,65%	15,92%
Total por Columna	49	1898	358	222	99	2626
	1,87%	72,28%	13,63%	8,45%	3,77%	100,00%

Tabla 11 Resumen estadístico para Número de idiomas según Indicador de mora

Sin embargo, estudiando la probabilidad condicionada de que, para un número de idiomas determinado, el cliente sea moroso tiene un comportamiento curioso: desciende a medida que aumentan los idiomas hasta llegar dos (como se da por hecho que los clientes hablan español, serán dos idiomas hablados además de la lengua materna), y en ese punto comienza a ascender la probabilidad de *default*.

P(1|1)	P(1|2)	P(1|3)	P(1|4)	P(1|+ de4)
0,1602	0,1480	0,1577	0,1717	0,1836

Tabla 12 Probabilidad de mora condicionada al número de idiomas

En el análisis del ámbito de operación de la empresa en la que trabaja el cliente según el indicador de mora, encontramos que, en nuestra muestra, a medida que las empresas están más extendidas, aumenta la probabilidad de mora. Así, un cliente que trabaja en una empresa internacional tiene mayor probabilidad de mora que un cliente que trabaja en una empresa nacional, el cual a su vez tiene mayor probabilidad de mora que un cliente que trabaja en una empresa local. Se detallan las probabilidades en la TABLA 14.

En primera instancia el sentido de negocio nos podría incitar a catalogar el análisis como erróneo, pero se debe tener en cuenta que la base datos de la que se parte fue creada durante los años de la crisis de 2008, por lo que el comportamiento de la probabilidad de mora no tendría por qué ser descabellado. En cualquier caso, la distribución de observaciones no muestra nada discriminante, por lo que se incluirá la variable al modelo para estudiar su predictibilidad.

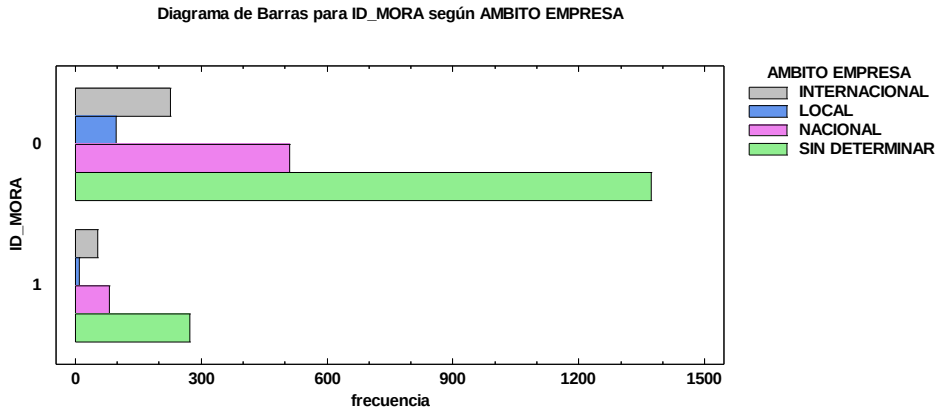

Figura 12 Ámbito de empresa según Indicador de mora

	INTERNACIONAL	LOCAL	NACIONAL	SIN DETERMINAR	Total por Fila
0	226	97	512	1373	2208
	8,61%	3,69%	19,50%	52,28%	84,08%
1	54	10	82	272	418
	2,06%	0,38%	3,12%	10,36%	15,92%
Total por Columna	280	107	594	1645	2626
	10,66%	4,07%	22,62%	62,64%	100,00%

Tabla 13 Resumen estadístico para Ámbito de empresa según Indicador de mora

P(1|Internacional)	P(1|Nacional)	P(1|Local)
0,1929	0,1380	0,0935

Tabla 14 Probabilidad de mora condicionada por el ámbito de la empresa

6.1.2 Análisis de variables continuas

Para el análisis de las variables continuas, se realizará un diagrama de cajas múltiple para cada una, para clientes sanos y morosos por separado, comparando los principales estadísticos en cada caso. Se realizará con procedimientos de Statgraphics.

En el análisis de la antigüedad laboral del cliente en función del tipo de cliente, se observa que la variabilidad de la antigüedad laboral en el grupo de morosos es mayor que en el grupo de clientes sanos.

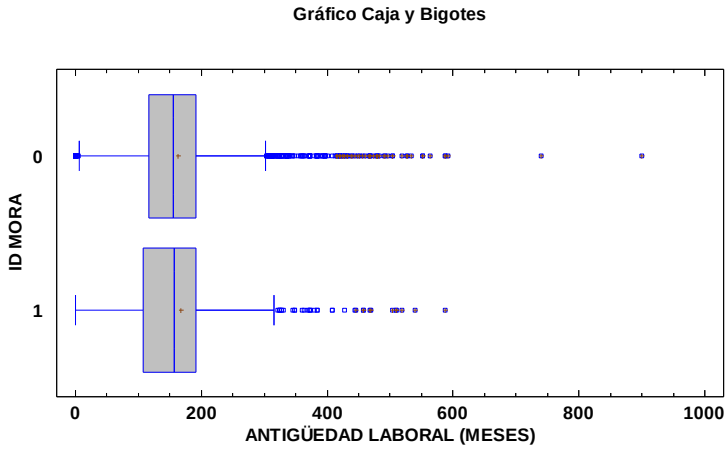

Figura 13 Antigüedad laboral según Indicador de mora

ID MORA	*Recuento*	*Promedio*	*Mediana*	*Desviación Estándar*	*Mínimo*	*Máximo*	*Rango*
0	2208	163,306	156,0	91,3028	0	900,0	900,0
1	418	167,536	157,0	101,432	0	588,0	588,0
Total	2626	163,979	156,0	92,9814	0	900,0	900,0

Tabla 15 Resumen Estadístico para Antigüedad laboral

Del análisis de la antigüedad en el puesto actual según el indicador de mora, se observa que los clientes morosos tienen, en mediana, menor antigüedad en el puesto. Además la variabilidad es mayor en el caso de los morosos.

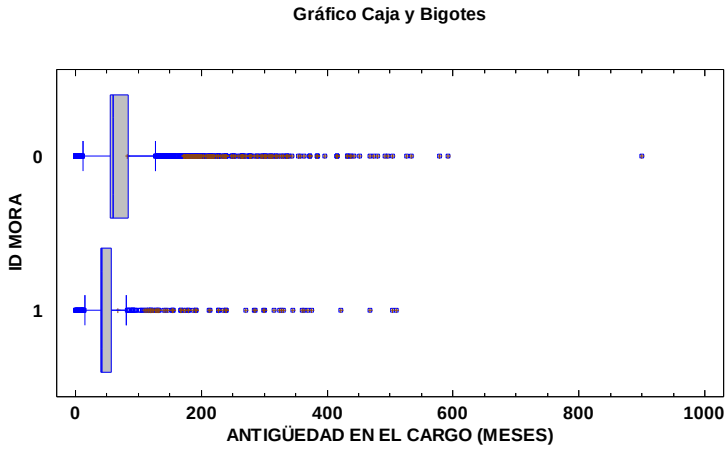

Figura 14 Antigüedad en el cargo según Indicador de mora

ID MORA	*Recuento*	*Promedio*	*Mediana*	*Desviación Estándar*	*Mínimo*	*Máximo*	*Rango*
0	2208	82,2219	60,0	76,9407	0	900,0	900,0
1	418	67,3254	42,0	77,5588	0	509,0	509,0
Total	2626	79,8507	60,0	77,2172	0	900,0	900,0

Tabla 16 Resumen estadístico para Antigüedad en el cargo

En el análisis del tiempo total estudiado por el cliente, incluyendo, si lo informa en la red social, la etapa anterior a la universidad, no se observa apenas diferencia entre el grupo de morosos de nuestra muestra y el grupo de clientes sanos. Llama la atención la gran concentración de datos entorno a 6 años, aunque tiene sentido, ya que la mayoría de las carreras universitarias terminaban en un tiempo entorno a los 5-6 años.

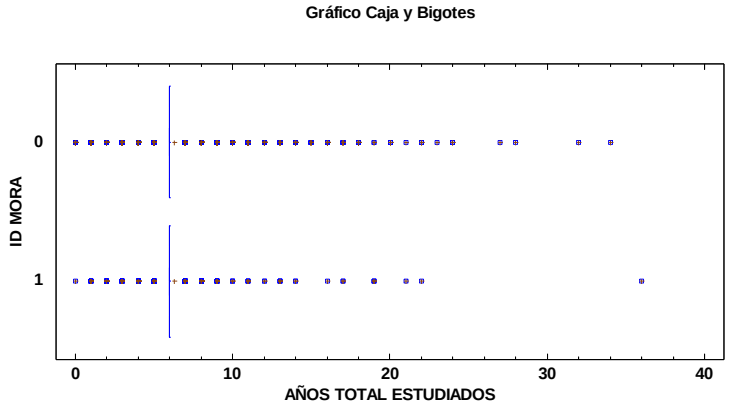

Figura 15 Número de años estudiados según indicador de mora

ID MORA	*Recuento*	*Promedio*	*Mediana*	*Mínimo*	*Máximo*	*Rango*
0	2208	6,32156	6,0	0	34,0	34,0
1	418	6,29665	6,0	0	36,0	36,0
Total	2626	6,31759	6,0	0	36,0	36,0

Tabla 17 Resumen estadístico para Número de años estudiados según indicador de mora

En el análisis del tiempo transcurrido desde la última vez que estudió el cliente según el indicador de mora, observamos que, en mediana, los clientes morosos han dejado transcurrir más tiempo desde la última vez que estudiaron. Vuelve a ser llamativa la gran concentración de datos entorno a 10 años en el caso de clientes sanos, y 12 en el caso de clientes morosos.

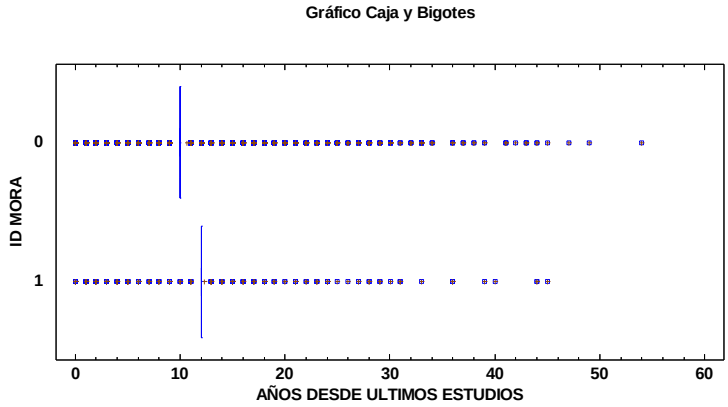

Figura 16 Tiempo desde los últimos estudios según indicador de mora

ID MORA	*Recuento*	*Promedio*	*Mediana*	*Mínimo*	*Máximo*	*Rango*
0	2208	10,6563	10,0	0	54,0	54,0
1	418	12,2727	12,0	0	45,0	45,0
Total	2626	10,9136	10,0	0	54,0	54,0

Tabla 18 Resumen estadístico para Tiempo desde los últimos estudios

Analizando la duración máxima en un cargo según el indicador de mora, no se observa una diferencia muy llamativa en los cuartiles, si bien es cierto que los clientes morosos tienen en mediana mayor duración máxima en los cargos.

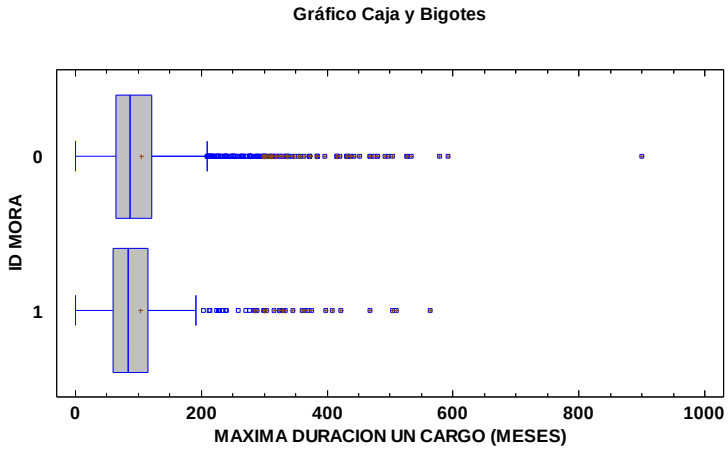

Figura 17 Máxima duración en un cargo según Indicador de mora

ID MORA	*Recuento*	*Promedio*	*Mediana*	*Desviación Estándar*	*Mínimo*	*Máximo*	*Rango*
0	2208	105,183	87,0	78,9726	0	900,0	900,0
1	418	102,957	84,0	84,3186	0	564,0	564,0
Total	2626	104,829	87,0	79,835	0	900,0	900,0

Tabla 19 Resumen estadístico para Máxima duración en un cargo

Analizando la variable *Duración mínima en un cargo* según el indicador de mora, encontramos en nuestra muestra que los clientes sanos tienen, en mediana, una duración mínima mayor que los clientes morosos. Además, en los clientes morosos hay una menor dispersión de los datos.

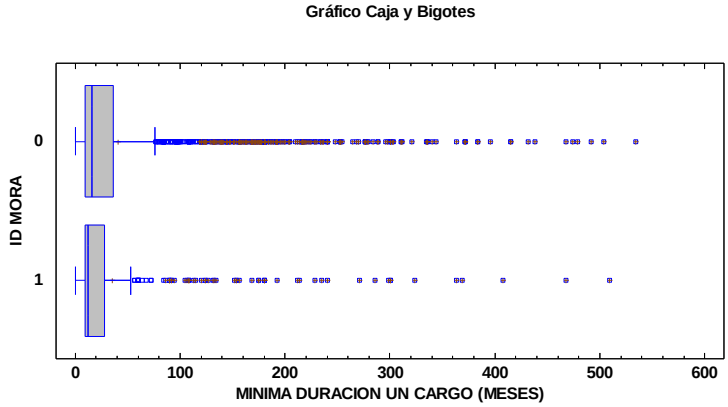

Figura 18 Mínima duración en un cargo según Indicador de mora

ID MORA	*Recuento*	*Promedio*	*Mediana*	*Desviación Estándar*	*Mínimo*	*Máximo*	*Rango*
0	2208	41,1132	16,0	67,9192	0	534,0	534,0
1	418	35,6411	12,0	67,6921	0	509,0	509,0
Total	2626	40,2422	16,0	67,8998	0	534,0	534,0

Tabla 20 Resumen estadístico para Mínima duración en un cargo

6.2 Análisis multivariante

El análisis multivariante implica estudiar todas las variables susceptibles de entrar en el modelo, a fin de descartar las que estén correladas o las que sean independientes de la morosidad. Se ha llevado a cabo un estudio del que se obtiene el resultado del Test de Chi-Cuadrado para las variables discretas, y una matriz de correlaciones para las variables continuas.

6.2.1 Test Chi-Cuadrado: Asociación de variables discretas

Para el estudio de las variables categóricas, se realizará el Test de asociación de la Chi-Cuadrado. Se toma como hipótesis nula H_0 la ausencia de asociación entre las variables.

H_0: No existe asociación entre las variables.
H_1: Existe asociación entre las variables.

En la Tabla 21 se resumen los resultados del Test Chi-Cuadrado. Como se puede observar, existe evidencia muestral para concluir que la variable *Ámbito de la empresa* tiene asociación significativa con la variable *Indicador de* mora.

Indicador de Mora Vs:	P-Valor
Nº trabajos históricos	0,4230
Sector de producción	0,1685
Categoría salarial	0,4806
Nº de idiomas hablados	0,9531
Ámbito de la empresa	0,0406

Tabla 21 Test de Chi-Cuadrado

6.2.2 Matriz de correlaciones

Para el análisis de la correlaciones entre las variables continuas, se construirá una matriz de correlaciones, que se muestra en la Tabla 22.

Como se observa en dicha tabla, hay varias variables cordeladas. Las variables *Duración máxima en un cargo* y *Duración mínima en un cargo* se encuentran muy correladas, aunque era de esperar dada la naturaleza de dichas variables. En muchos casos es posible que el valor coincida. Sucede lo mismo entre éstas y las variables *Antigüedad en el puesto actual* y *Antigüedad Laboral*. Si el cliente de nuestra muestra llevara poco tiempo trabajando y sólo ha tenido un puesto, el valor sería el mismo para las cuatro variables antes mencionadas.

Sin embargo, no es tan latente la correlación entre *Antigüedad Laboral* y *Antigüedad en el puesto actual*, y las variables relacionadas con los estudios no están correladas con las variables relacionadas con el ámbito profesional.

	Tiempo total estudiado	Tiempo desde la última vez que estudió	Antigüedad Laboral	Antigüedad en el puesto actual	Duración máxima en un cargo	Duración mínima en un cargo
Tiempo total estudiado		-0,1131	0,0487	0,0487	0,0249	-0,0402
Tiempo desde la última vez que estudió	-0,1131		0,3610	0,2698	0,3228	0,1704
Antigüedad Laboral	0,0470	0,3610		0,5691	0,8137	0,3426
Antigüedad en el puesto actual	0,0487	0,2698	0,5691		0,8189	0,6167
Duración máxima en un cargo	0,0249	0,3228	0,8137	0,8189		0,6391
Duración mínima en un cargo	-0,0402	0,1704	0,3426	0,6167	0,6391	

Tabla 22 Matriz de correlaciones.

Se decide no incluir las variables *Duración máxima en un cargo* y *Duración mínima en un cargo* en el modelo por estar altamente correladas.

6.3 Modelo de regresión seleccionado

En la construcción del modelo con el programa R, se ha dividido la muestra en dos partes: una muestra de training, que supone el 60% de la muestra total y alimentará el modelo, y una muestra de test, que supone el 40% restante y será utilizada para estudiar la bondad de ajuste del modelo. Dichas muestras contienen observaciones de morosos y sanos por partes iguales, y han sido elegidas de forma aleatoria.

A continuación, se muestra la tabla 23 con los datos del modelo ajustado.

Variable	Coeficiente	Std. Error	P-Valor	Significación	Odds Ratio
Constante	-1,5252	0,2244	1,08E-11	***	0,22
Tiempo (años) desde el último año de estudio (x_1)	0,0576	0,0101	1,62E-08	***	1,06
Antigüedad (meses) en el puesto actual (x_2)	-0,0053	0,0012	2,62E-05	***	0,99
Empresa de ámbito 'Local' (x_3)	-0,9729	0,4768	0,0413	*	0,38
Empresa de ámbito 'Nacional' (x_4)	-0,5016	0,2483	0,0434	*	0,61
Empresa de ámbito 'Sin_determinar' (x_5)	-0,4164	0,2129	0,0505	.	0,66

Código de significación:	0 '***' 0,001 '**' 0,01 '*' '0,05' . ' 0,1 ' '

Tabla 23 Modelo de predicción de mora

Como se puede observar en la tabla, todas las variables tienen un nivel de significación muy elevado, a excepción del caso en el que no se especifique el ámbito de la empresa en la que trabaja el cliente. El modelo quedaría escrito como:

$$Logit\left(\frac{\pi_i}{1-\pi_i}\right) = -1{,}5252 + 0{,}0576x_1 - 0{,}0053x_2 - 0{,}9729x_3 - 0{,}5016x_4 - 0{,}4164x_5$$

Se estudiará la bondad de ajuste del modelo mediante la Tasa de Clasificaciones Correctas y el área de la curva ROC, tal como se especificó en la sección 5.3. Se han calculado ambas medidas con un punto de corte igual a 0.16, siendo este valor la proporción de morosos en la muestra, ya que es el valor que maximiza la Tasa de Clasificaciones Correctas (Hosmer, Lemeshow, & Sturdivant, 2013). En la tabla 24 se pueden ver la Tasa de Clasificaciones Correctas y la medida del área bajo la curva ROC, y en la figura 19 la gráfica que representa la curva de ROC, la cual enfrenta sensibilidad frente a especificidad. Con estas dos medidas podemos concluir que el modelo califica bien más del 70% de los casos.

Tasa de clasificaciones correctas	Área bajo la curva ROC
72,95%	0,72

Tabla 24 TCC y área bajo la curva ROC

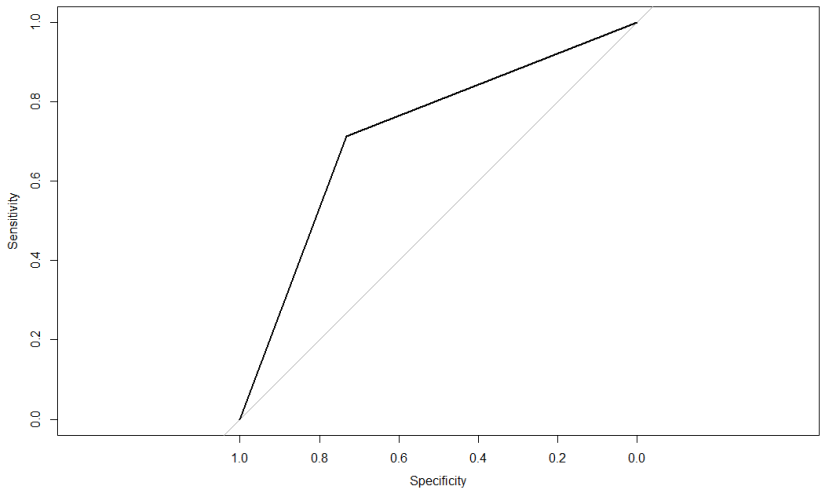

Figura 19 Curva ROC del modelo

El modelo debe interpretarse en términos de la Odds Ratio, como se especificó en la sección 5.6.

Así, por ejemplo, si el tiempo transcurrido desde el último año de estudio se incrementa en 12 años, la ventaja a favor de ser moroso se duplica.

En la misma línea, si la antigüedad en el puesto actual aumenta en 132 meses (11 años) la ventaja a favor de ser moroso se ve reducida a la mitad.

Para el caso de la variable *Ámbito de la empresa*, al tratarse de una variable categórica, se crea una variable *dummie* para cada categoría, dejando una categoría como referencia. Vemos que sólo son significativas las categorías Local y Nacional, dejando Internacional como categoría de referencia. Así, considerando la inversa de 0.38 (Odd Ratio para la categoría Local), la ventaja a favor de ser moroso es aproximadamente 2.63 veces mayor si la empresa es Internacional que si es de ámbito local. Por otro lado, considerando la inversa de 0.61 (Odd Ratio para la categoría Nacional), la ventaja a favor de ser moroso es aproximadamente 1.6 veces mayor si la empresa es de ámbito internacional que si es de ámbito nacional.

7 Conclusiones y líneas abiertas

En este Trabajo de Fin de Grado se pretendía cuantificar el valor de la información contenida en redes sociales con un caso práctico: un modelo de Scoring.

Para ello, se han estudiado las diferentes redes sociales atendiendo a su penetración en el territorio español y al tipo de información que contenían. Posteriormente se ha experimentado sobre la forma de extraer dicha información, desarrollando diferentes programas que permitían almacenar dicha información.

Una vez capturada la información, es necesario un laborioso trabajo de limpieza y tratamiento de datos, cuyo paso más fatigoso es la normalización de los mismos. Sobre todo ello, el problema más agudo es la desambiguación de perfiles de manera inequívoca.

Con los datos correctamente tratados, se construye una tabla que contiene toda la información extraída de internet y la información facilitada por una entidad financiera. Sobre esta base de datos se construye un modelo de regresión logística, resultando una Tasa de Clasificaciones Correctas del 72,95%.

En el modelo se han incluido variables relacionadas con la información de los estudios y el trabajo, siendo las más predictivas el número de años que han transcurrido desde el último año en que se estudió, la antigüedad del cliente en el puesto en el que trabaja actualmente, y el ámbito de operación de la empresa para la que trabaja el cliente (Local, Nacional o Internacional). Dichas variables no existen en las bases de datos bancarias tradicionales, por lo que conjugar esa información con la ya existente podría elevar el poder predictivo de los modelos sustancialmente.

Por otro lado, la captura y el tratamiento de información precisan de una metodología más automática y con mayor nivel de precisión para la construcción de grandes bases de datos de redes sociales. Mediante la captura ordenada de datos por parte de las entidades financieras será posible un mayor aprovechamiento de la información.

En definitiva, se puede concluir que las redes sociales contienen información que puede ser de interés para la gestión de riesgos en una entidad financiera, y dado que las nuevas tecnologías permiten capturar y tratar dicha información más sencilla y masivamente, se ha propuesto un proyecto de aplicación real sobre las bases de este Trabajo de Fin de Grado.

Para terminar, se ha programado una herramienta que simplifica el uso del modelo, poniéndolo al alcance de cualquier usuario sin necesitar conocimientos estadísticos ni financieros avanzados.

Bibliografía

1. adigital. (s.f.). *http://www.adigital.org/*. Obtenido de http://www.adigital.org/noticias/mas-de-la-mitad-de-las-empresas-afirman-que-el-retorno-en-redes-sociales-es-igual-o#sthash.04hDwdDJ.dpuf
2. Agresti, A. (2002). *Categorical Data Analysis* (Second ed.). Gainesville, Florida: Wiley.
3. Altman, E. I. (2001). *Managing credit risk: a challenge for the new millenium.* New York: NYU Stern School of Business.
4. Aluja, T. (2001). *La minería de datos, entre la estadística y la inteligencia artificial* (Vol. 25).
5. Arcarons Bullich, J., & Colange Ramírez, S. (2008). *Microeconometría. Introducción y aplicaciones con software econométrico para excel.* Delta Publicaciones.
6. Bernués Pardo, J., Lozano Imízcoz, M. T., & Polo Blanco, I. (2012). Selección de modelos matemáticos de la Universidad de Zaragoza. *Universidad de Zaragoza , 15*, 187-204.
7. Cogneau, P., & Zakamouline, V. (2010). *Bootstrap Methods for Finance: Review and Analysis.*
8. Couso, A. S. (2011). *Modelos de respuesta discreta en R y aplicación con datos reales.* Universidad de Granada.
9. Eckerson, W. (2002). Data Warehousing Special Report: Data quality and the bottom line.
10. Fed. (2012). Supervisory Guidance on Model Risk Management. *Board of Governos of the Federal Reserve System* .
11. Harrel, F. E. (2002). *Regression Modeling Strategies* (1st ed.). Springer.
12. Hernández-Orallo, J., Ferri, C., Lachinche, N., & Flach, P. (2004). *Roc Analysis in Artificial Intelligence.* Valencia: ROCAI.
13. Hosmer, D., Lemeshow, S., & Sturdivant, D. (2013). *Applied Logistic Regression* (3 ed.). Wiley.

14. Huang, Z. C.-J.-H. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study, Decision support systems.
15. IAB. (s.f.). *IAB Spain*. Obtenido de http://www.iabspain.net/
16. INE. (2013). Decil de salarios del empleo principal. Encuesta de Población Activa (EPA).
17. Kovalerchuk, B. (2009). *Data Mining for Financial Applications.*
18. Lai, T. L. (2010). Data Science, Statistical Modeling, and Financial and Health Care Reforms. *Department of Statistics, Stanford University* , 1-18.
19. Llaugel, F. A., & Fernández, A. I. (2001). Evaluación del uso de modelos de regresión logística para el diagnóstico de instituciones financieras. *Ciencia y sociedad , XXXVI* (4), 590-627.
20. Loofbourrow, J. L. (1995). What AI brings to trading and portfolio management. *Artificial Intelligence in the capital markets* , 3-28.
21. Magoulas, R. (2011). Why the financial world should care about Big Data and Data Science.
22. Management Solutions. (2014). Model Risk Management.
23. Marr, B. (2015). *Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance* (1 ed.). Wiley.
24. MIT. (2005).
25. Mures Quintana, J. M., García Gallego, A., & Vallejo Pascual, E. M. (2005). *Aplicación del análisis discriminante y regresión logística en el estudio de la morosidad en las entidades financieras.* Castilla y León: Pecvnia.
26. OCC. (2011-12). Supervisory Guidance on Model Risk Management. *Office of the Comptroller of the Currency* .
27. Russel, M. A. (2013). *Mining the social web* (Second ed.). O'Reilly.
28. *SAS*. (s.f.). Obtenido de http://www.sas.com/es_mex/customers/local/bankinter-riesgos.html
29. *SAS*. (s.f.). Obtenido de http://www.sas.com
30. Siddiqqi, N. (2006). *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring.* New Jersey: Wiley.

31. Tussell, F. (Octubre 2011). Análisis de regresión. Introducción Teórica y Práctica basada en R.
32. Walzack, S. (2001). An empirical analysis of data requierements for financial forecasting with neural networks. *Journal of Management Information Systems , 17(4)*, 203-222.
33. Zhang, L. (2013). Sentiment Analysis on Twitter with Stock Price and Significant keyword Correlation. *Department of Computer Science, The University of Texas at Austin* .

Anexo I: Construcción de un árbol de decisión

Durante la realización de este Trabajo se desarrolló un estudio paralelo, en el cual se evaluó la predictibilidad de los datos mediante un árbol de decisión.

Los árboles de decisión permiten asignar, a grupos predefinidos, toda la muestra de estudio, en función de una serie de variables predictivas. Teniendo variables de respuesta categórica, crean reglas simples mediante las que se subdivide la muestra. Se trata de una técnica discriminante, que fracciona la muestra en cada rama del árbol. Una vez se ha dividido una parte de la muestra, sólo podrá volver a dividirse una parte resultante de la primera división (véanse figuras 21 y 22).

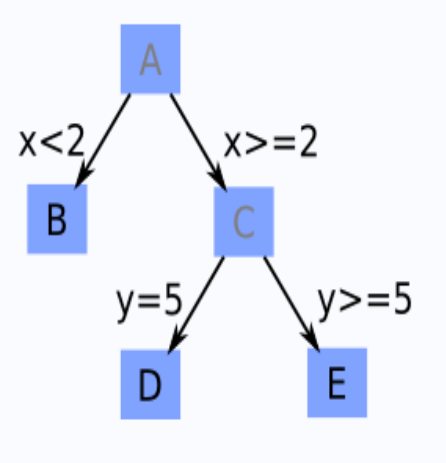

Figura 21 Esquema de un árbol de decisión

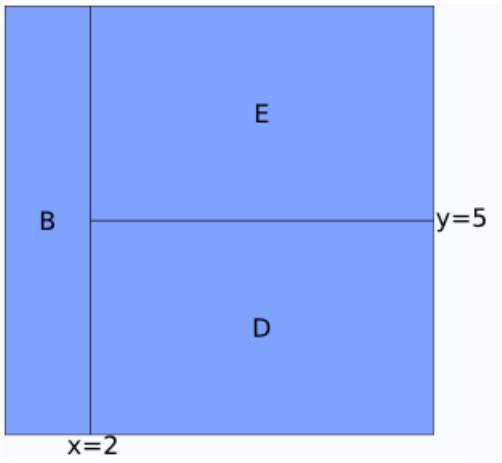

Figura 20 División de la muestra

Aunque el poder predictivo de un árbol es limitado, en el sector financiero se conjugan con modelos logit para hacer uso de ellos. Tienen a su favor puntos muy a tener en cuenta, como que las reglas de asignación son rápidas y simples (y por tanto fácilmente interpretables), o que se trata de una metodología válida para cualquier variable de entrada (ya sea continua, discreta o categórica).

Para la creación de un árbol de decisión se ha utilizado un procedimiento de SAS llamado HPSPLIT. Dicho código se genera un árbol de decisión, pudiendo especificar los siguientes criterios:

- Número mínimo de observaciones por hoja
- Número máximo de hojas por nodo
- Profundidad máxima del árbol
- Número de observaciones por nivel, para que se pueda considerar una división en él.
- P-valor máximo para considerar una división
- Ajuste de Bonferroni a los P-valores de después de la división
- Mínima distancia de Kolmogorov-Smirnov

Se realizó un modelo con 4 observaciones en el nodo final, otro con 6 y un último con 10 observaciones para estudiar la robustez del árbol.

Una vez realizado el árbol, SAS permite realizar una poda mediante la cual se eliminan los nodos con menos observaciones. Las observaciones de los nodos podados, se añaden al nodo anterior, con lo que se obtiene un árbol más consistente en cuando a número de observaciones por nodo.

Como resultado de dichos estudios, se resume el poder predictivo de los árboles en la Tabla 26.

Poder predictivo del árbol de decisión según los parámetros de entrada

Resultados Árboles	Nº Obs. Nodo Final	ROC
Con Poda	4	77,39%
	6	76,21%
	10	71,98%
Sin Poda	4	80,42%
	6	79,32%
	10	76,13%

Tabla 25 ROC de los distintos modelos

A pesar de tener una ROC relativamente elevada, la construcción de un árbol de decisión no supone una metodología excesivamente técnica, y por supuesto menos precisa que una regresión logística. Por ello, aunque deja abierta la puerta a futuras investigaciones, no puede ser tratado como un éxito en nuestro ejercicio. Sin embargo, incita a realizar una extracción mayor y un tratamiento más exhaustivo de los datos.

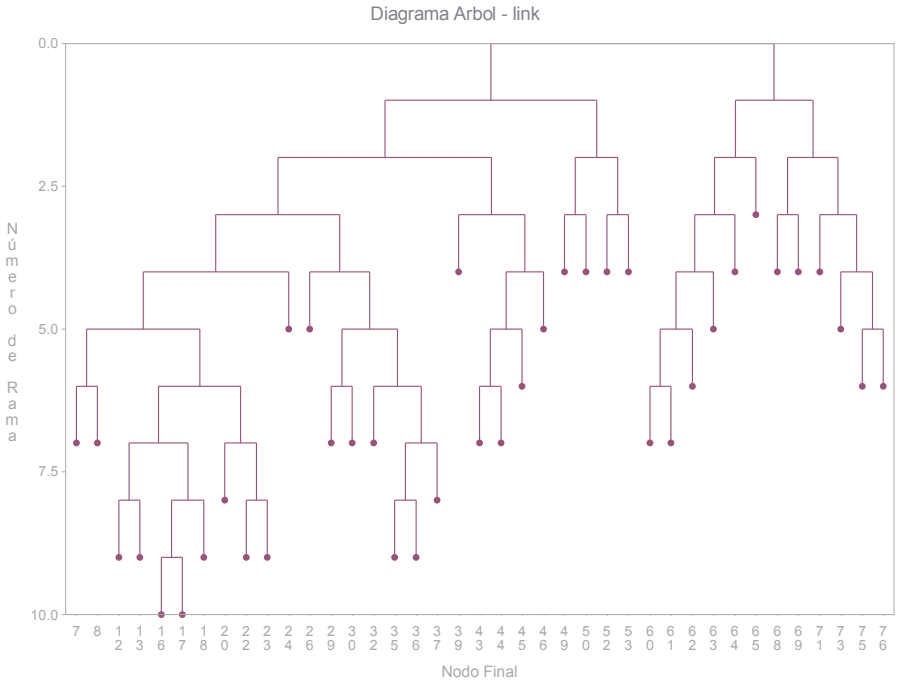

Figura 22 Árbol de decisión (podado) creado mediante SAS